Experimental and Theoretical Analysis of Multivesicular Release at a Sensory Synapse

Ben James

Submitted for the degree of Doctor of Philosophy

University of Sussex

February 2020

Declaration

Portions of this work (Chapters 3 and 4) have been published in *Nature Neuroscience* under the title “An amplitude code transmits information at a visual synapse”. For the portions included in this thesis, my contributions include: writing software, conceiving and designing experiments, performing initial pilot experiments, and writing the manuscript. I have attempted to restrict this thesis to work I primarily performed - although many experimental recordings were collected by co-authors, and I took no part in the establishment or maintenance of transgenic zebrafish lines.

Chapter 5: Theoretical Frameworks for the Analysis of MVR, was written and developed entirely by me, with a portion of the programming done by a student under my guidance.

I hereby declare that this thesis has not and will not be, submitted in whole or in part to another University for the award of any other degree.

Signature……

Summary

Traditionally, neurons were believed to transmit information between one another by action potentials and the release of synaptic vesicles, both of which thought to be binary processes. In this framework, an action potential (or change in membrane potential for graded neurons) elicits the release of at most a single vesicle per active zone (AZ), the site of vesicular release. Information, then, is conveyed in this system simply be altering the probability of a single vesicle being released, and thus altering the total rate of release.

In recent years, however, evidence has demonstrated that many cells, particularly those in the early sensory systems, are capable of multivesicular release (MVR), wherein multiple synaptic vesicles are released nearly simultaneously, with precision of up to 10 microseconds. As most neurons have a temporal integration window on the millisecond range, it is thus likely the postsynaptic cell could represent these events as distinct symbols in a non-binary digital system, allowing for information transmission not only in the rate of vesicular events, but also their amplitude. While evidence for MVR has existed for decades, considerably less attention has been paid to its functional significance – how does MVR affect processing in intact circuits? Is it utilized to transmit information? How efficiently does it operate?

In this work I attempt to answer these questions, first by developing a method with which one can decompose 2p glutamate recordings from intact zebrafish BC terminals into units of individual vesicles – essentially ‘counting’ vesicles. In doing so, I show that MVR is used in the early visual system to transmit temporal contrast information, one of the most basic aspects of the visual scene. Finally, I create a model of vesicle release and postsynaptic activity that allows for directly comparing how information transmission can be influenced by either rate or amplitude coding. Here, by systematically altering the parameters of the postsynaptic cell, I demonstrate in which circumstances either rate coding or amplitude coding are beneficial for transmission. MVR, rather than being some obscure phenomenon, plays a fundamental part in processing and transmitting information in early sensory systems.

Acknowledgements

I would like to thank first and foremost my supervisor Leon, as well as Sofie, Jose, and Lea – without the work of which this thesis would have suffered. I would additionally like to thank Paul Pichler for advice – and graduating before me. Lastly, I would like to acknowledge Ludwig Boltzmann, whose work inspired a great deal of the foundations of this work, and whose exit strategy the writing of this thesis forced me to consider.

$$\int \frac{1}{cabin} d(cabin) = log(cabin) + C = houseboat$$

-Thomas Pynchon, *Gravity's Rainbow*

Abbreviations

Neuroscientific

RGC: retinal ganglion cell, BC: bipolar cell, AC: amacrine cell, HC: horizontal cell, PR: photoreceptor, LIF: leaky integrate-and-fire, RRP: readily releasable pool, RP: reserve pool, CNS: central nervous system, AZ: active zone

Mathematical

PP: Poisson Process, HPP: Homogeneous Poisson Process, NHPP: Nonhomogeneous Poisson Process, pdf: probability density function, pmf: probability mass function, cdf: cumulative distribution function, mgf: moment generating function, LOTUS: law of the unconscious statistician, GMM: Gaussian Mixture Model, HMM: Hidden Markov Model, IMM: Infinite Mixture Model

Table of Contents

Table of Figures

Chapter 1: General Introduction

1.1: The Fundamental Goal of Sensory Systems, Historical Perspectives

As you read this, the approximately 100 billion neurons in your nervous system periodically activate (Kandel, Schwartz et al. 2000). Information in the form of light (or perhaps the sounds of a car heard idly passing by through the window, or even the taste of the coffee you absentmindedly sip) is detected by sensory neurons. This information is then passed through a complex series of circuits, and finally fully processed in the central nervous system, resulting in the perception of words (or automotive noise, or bitter-sweet caffeinated gustatory delight). The question of how (and even where) these sensations arise has interested researchers for centuries. At first confined to the purely philosophical caves and shadows of Plato, the advent of experimentation has greatly advanced our understanding of this process, and the underlying biological processes that give rise to it.

While the question of exactly *how* a percept consciously arises within the nervous system is still largely an open question, the manner in which early sensory systems operate is shrouded in significantly less mystery. Though the precise mechanisms with which each sensory system operates can be quite diverse, a degree of consensus has been reached regarding the overarching or fundamental goal of sensory systems: to maximize information on relevant sensory stimuli while simultaneously minimizing metabolic cost (MacKay and McCulloch 1952, Barlow 1961, Balasubramanian, Kimber et al. 2001, Laughlin 2001, Schwartz and Simoncelli 2001), first propped by Horace Barlow in 1961 as the Efficient Coding Hypothesis. Here, sensory systems are thought to be optimized to transmit useful information (such as the predatory information involved in a frog catching a fly (Lettvin, Maturana et al. 1959)), with minimal energetic cost. Many might consider the fundamental question of sensory neuroscience, then, as “how does the nervous system accomplish this feat?” In what ways does the early sensory system deal with this information/energy tradeoff, and how is information processed in early sensory systems? Specifically, how does early sensory information become integrated into the nervous system to an organism’s advantage? While a full review on the

history of how science has attempted to illuminate this answer goes beyond the scope of this work, considerable insight can be gained from the historical perspective – how have past researchers attempted to answer this question?

While some understanding on the nervous system was gained throughout a large portion of human history, it wasn't until the 20th century when true steps in the advancement of modern neuroscience took place. Here, a series of 'call-and-response' question and answer pairs pushed understanding, gradually building and refining the picture of neural operation. First, there arose the question of the nature of neurons. In what is known as the continuous or contiguous debate, researchers questioned the nature of the neuron. Are neurons distinct and separate cells, operating as individual units, or do they form one continuous process? In what many consider the first great achievement of modern neuroscience, Ramon y Cajal definitively proved the former – that neurons are distinct units, the smallest biological unit in the nervous system. This question being answered, the next arose – how do these fundamental units communicate with one another? In another Nobel Prize winning work, Otto Loewe, supposedly inspired by a dream, answered this question by identifying the first neurotransmitter – acetylcholine – and thus demonstrated that neurons communicate via the release of chemical neurotransmitter (Karczmar 1996). Near the same time, Bernard Katz further illuminated the nature of this chemical communication – that neurotransmitter is not continuous but rather quantized, with a fundamental quantal unit of neurotransmitter (Del Castillo and Katz 1954). Hodgkin and Huxley then showed the electrophysiological nature of the action potential – transmitter is released following a regenerative electrical signal in the axon of neurons, and the electrical properties of these symbols can be well described by the dynamics of current for sodium and potassium ions (Hodgkin and Huxley 1952).

Although the discoveries mentioned above are a miniscule portion of the discoveries in neuroscience in the past century, it is beneficial here to pause and discuss how these advancements can build upon the answer to the fundamental goal of neuroscience. Neurons – distinct biological units – communicate to one another by the quantized release of synaptic chemical neurotransmitter. In a coarse manner, then, we have already answered a chunk of the question. Using this knowledge, we can begin to understand how information is represented and transmitted in neural systems.

1.2: The Spike, The Vesicle, and Rate Coding

The action potential – the ubiquitous neural spike – has for centuries been paired with the synaptic vesicle as one of the two most fundamental units of neural information representation and transmission. Filling pages of virtually every introductory book on neuroscience, the spike has been described in detail: its electrochemical basis, its link with learning and long-term potentiation, and more importantly for this work – its binary 'all-or-nothing' nature and consequent use in information transmission. The biochemical processes involved in action potential are well-studied. In brief, neurons can be considered to function as leaky integrators – this formulation in fact has existed for over a century (Abbott 1999). Current flows into the cell, most often as a result of the binding of synaptic neurotransmitter and consequential opening of ion channels, generating a difference in voltage across the cells membrane. In the absence of a spike, this sub-threshold activity gradually decays due to leaky ion channels. However, if the membrane potential reaches a threshold, commonly denoted θ, voltage-gated ion channels are opened. This results in a regenerative signal – positive ions flow inwards, further depolarizing the cell and activating more voltage-gated ion channels – carried across the axon and terminating at the axon terminal, usually resulting in the release of synaptic neurotransmitter. Because these action potentials are not graded, with each cell exhibiting a stereotypical spike shape, spikes are said to be binary – there exists no 'double' or 'half' spike in the neural realm, simply the presence (one) or absence (zero) of a spike.

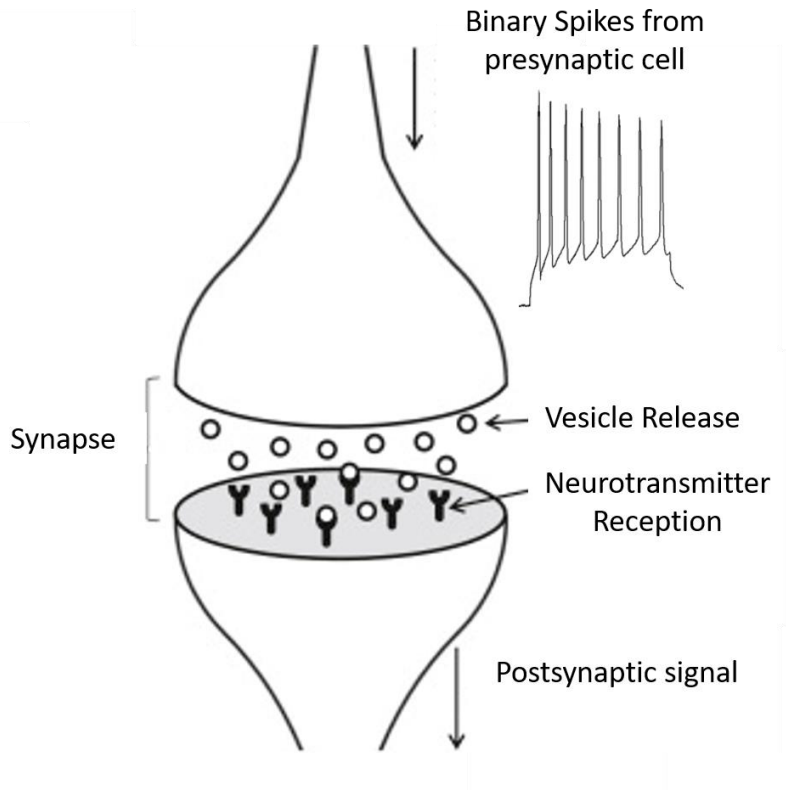

Figure 1.1: Basic picture of the spiking synapse.

Spikes travel down the presynaptic cells axon (top), resulting in the release of synaptic neurotransmitter. This neurotransmitter then diffuses across the synaptic cleft, binding to postsynaptic receptors and opening ion channels, resulting in the activation of the postsynaptic cell. Adapted from (Widrow, Kim et al. 2019).

In nearly every area of the nervous system studied the spike has been found and implicated in information processing. Place cells of the vertebrate hippocampus carry information pertaining to an organisms location in their spikes (O'Keefe 1976), while retinal ganglion cells spike in response to complex visual properties (Masland 2001, Gollisch and Meister 2008, Gollisch and Meister 2010, Baden, Nikolaev et al. 2014). Spiking activity can even be used to identify an animal's behavioral state or induce behavioral responses (Calhoun, Pillow et al. 2019), and activity of certain cells can inhibit or activate an organisms behavior ranging from avoidance responses (Temizer, Donovan et al. 2015) to predatory actions (Han, Tellez et al. 2017). The neural action potential, however, does not in itself convey information to other neurons (other than in the rare instances of electrical connections) – it is through the release of synaptic transmitter that one neuron 'communicates' to another. Luckily enough, as vesicular release at most synapses appeared to be binary, with a single release site capable of releasing

at most a single vesicle at one (Redman 1990, Korn, Sur et al. 1994, Yusim, Parnas et al. 2001), vesicle release can be studied in the same fashion as spiking activity, albeit with a slightly more complicated aspect relating to the increased stochasticity involved in quantal failures (Levy and Baxter 2002).

Even the cells that do not show spiking behavior, such as many cells in the early sensory systems, and rather encode information in a continuous or graded manner by modulating their membrane potentials are still thought to convey information solely by the modulation rate of release of vesicular neurotransmitter (de Ruyter van Steveninck and S.B. 1996, Juusola, French et al. 1996), excluding the rarer cases of gap junctional connections. Consequently, most theories of neural information transmission hypothesize information is represented by binary spiking/vesicle release processes, where all information is contained in the rate or timing of a symbol of fixed amplitude. In response to changes in input stimuli, cells of this type can represent information by either simply altering the frequency or timing of spikes or vesicle release. Notably, either of these strategies are highly dependent upon brain region, and likely reflect relative amounts of noise carried throughout the system, as well as the importance of each type of signal being transmitted. Early sensory systems, which must transmit signals with high temporal precision (Berry, Warland et al. 1997), tend to contain more information in precise spike timing than higher cortical areas, which are characterized by more irregular spike times (Shadlen and Newsome 1998).

Notably, as either graded or binary cells were thought to transmit information via modulating the release of a fixed number of vesicles, information between these cells was believed to consist of a **rate code** (Sejnowski 1995, Gerstner, Kreiter et al. 1997, Gautrais and Thorpe 1998, Borst and Theunissen 1999, Huxter, Burgess et al. 2003, London, Roth et al. 2010). In defining what I refer to as rate coding, note that I have chosen to broaden the definition of rate coding beyond the scope it is originally used. Many define neural rate code alongside neural time coding. Rate coding in those situations applies simply to information transmitted solely by modulating the rate of events regardless of the precise time of each event, which is used to define timing information. Here, I define rate coding to encompass both these scenarios – both information contained in rate and timing. The fundamental aspect which defines rate coding in the context of this work is that it operates via a binary signal consisting of either the presence or absence of a neural event, as opposed to a set of multiple symbols, which we will explore later. In this context, I describe rate coding not in the typical sense, but strictly in opposition of **amplitude coding**, wherein multiple synaptic vesicles worth of

neurotransmitter are released simultaneously, thus increasing the amplitude of the synaptic event, but not the rate of synaptic events. First, it is beneficial to describe the architecture of the visual system – an ideal system to explore aspects of information transmission in the nervous system.

1.3: Early sensory systems, the retina, and bottlenecks

Of all systems in the nervous system, we are best equipped to understand the early sensory systems (vision, audition, electroreception, etc.), for a large part due to the relatively simple input-output architecture compared to higher cortical areas. For example: the retina, the first stage in visual processing and one of the better-understood areas of the brain, operates largely in a uni-directional fashion (Masland 2001), see **Fig1.2** for a model of the excitatory pathway of the vertebrate retina: photoreceptors -> bipolar cells -> ganglion cells -> brain. Visual processing begins in the photoreceptor layer, where isomerization of opsins in rod and cone photoreceptors (PRs) results in a decrease in the release of synaptic glutamate in response to increases in light intensity. This information is then passed to the Bipolar Cells (BCs) of the outer nuclear layer, where it is split into ON and OFF channels (responsive to light onset or offset) by the presence of either sign-inverting metabotropic glutamate receptors (for ON cells), or sign-conserving ionotropic glutamate receptors (for OFF cells). Inhibitory horizontal cells (HCs) of this layer also function to further sharpen and tune the visual signal, before BCs release synaptic glutamate upon retinal ganglion cells (RGCs) and amacrine cells (ACs). ACs, inhibitory interneurons like HCs, further refine the visual signal, while RGC send all visual information to the brain through the optic nerve in a series of action potentials.

Processing within the retina thus follows a very straight-forward pathway where input (light) can be easily modulated experimentally, and the visual signal can be ‘followed’ by measuring the response of each cell type to various forms of visual stimuli. Photoreceptors, the first cells involved in image-forming vision, operate in a graded manner. In the dark state – high concentrations of cGMP allow sodium channels in the cells membrane to remain open, causing a net influx of positive ions and an overall depolarized state. As the light levels increase, photons begin to strike the PRs and are captured by light-sensitive molecules known as opsins. Photon absorption induces isomerization of these opsins, activating a chemical cascade which results in the reduction of cGMP in the cell, closure of sodium channels, and cell hyperpolarization that is scaled with the magnitude of the light intensity. PRs then are capable of encoding light intensity, which at high enough levels can be considered continuous, by modulating their membrane polarization, another virtually continuous metric. This information

can then be reliably synaptically relayed to postsynaptic cells, at the cost of increased release rates. Additionally, inhibitory action of Horizontal Cell interneurons (HCs) provides a further refinement of the visual signal at this stage, important for refinements of higher order neural responses.

Following along the stream of visual information, we next find the bipolar cells (BCs), one of the synaptic recipients of glutamatergic inputs of the PRs. Like the previous cells, we also find that these cells seem to operate largely in a graded manner, although the types of information they encode are considerably more sophisticated than the PRs. Rather than simply responding to light intensity, BCs can modulate their membrane potentials in a more complex fashion, responding to such properties of the visual scene as contrast (Odermatt, Nikolaev et al. 2012) or orientation (Antinucci, Suleyman et al. 2016). BCs are even capable of producing a predictive code, where stimulus changes, rather than the raw stimulus itself, are encoded (Johnston, Seibel et al. 2019). Notably, the visual signal largely remains a linear function of light intensity in BCs until the level of the BC output, the axon terminal, where further refinement occurs with inhibitory ACs acting on the BC terminal in a with a variety of inhibitory feedback mechanisms (Demb, Zaghloul et al. 2001): BCs release excitatory glutamate onto RGCs and ACs, and the ACs then feedback to BC terminals, releasing inhibitory GABA or glycine. This local inhibition allows for specific computations at the level of the BC axon terminal, allowing single cells to transmit multiple types of information (Nikolaev, Leung et al. 2013).

The final excitatory cell type in the retina is the RGC. Unlike PRs and BCs, RGCs operate on a purely binary framework – rather than continuously modulating their membrane potential, they carry information in sequences of action potentials. Processing in the retina, then, roughly follows three distinct phases – initial light detection by PRs, synaptic transmission through the BCs to the inner plexiform layer, and final processing and output of visual information from the optic nerve. Notably, the optic nerve in this system represents a bottleneck, in multiple ways. For one, there are a finite number of RGCs that can spike at a finite rate, limiting the total information transmission capabilities. Secondly, PRs generally outnumber RGCs so simply relaying what the PRs detect is also impossible. For instance, outside the monkey fovea, the cones outnumber RGCs by up to 50 times at higher eccentricities (Wassle and Boycott 1991). It is thus understandable that a great deal of information processing occurs prior to the optic nerve – as the retina cannot transmit all visual information to the brain, important information should preferentially be transmitted through the optic nerve, while extraneous or redundant information should be discarded. One way in which the retina

accomplishes this goal is by extracting sets of visual features (such as orientation, contrast, or even motion) from the visual scene before transmitting the information to the brain – in essence performing small computations of the visual information.

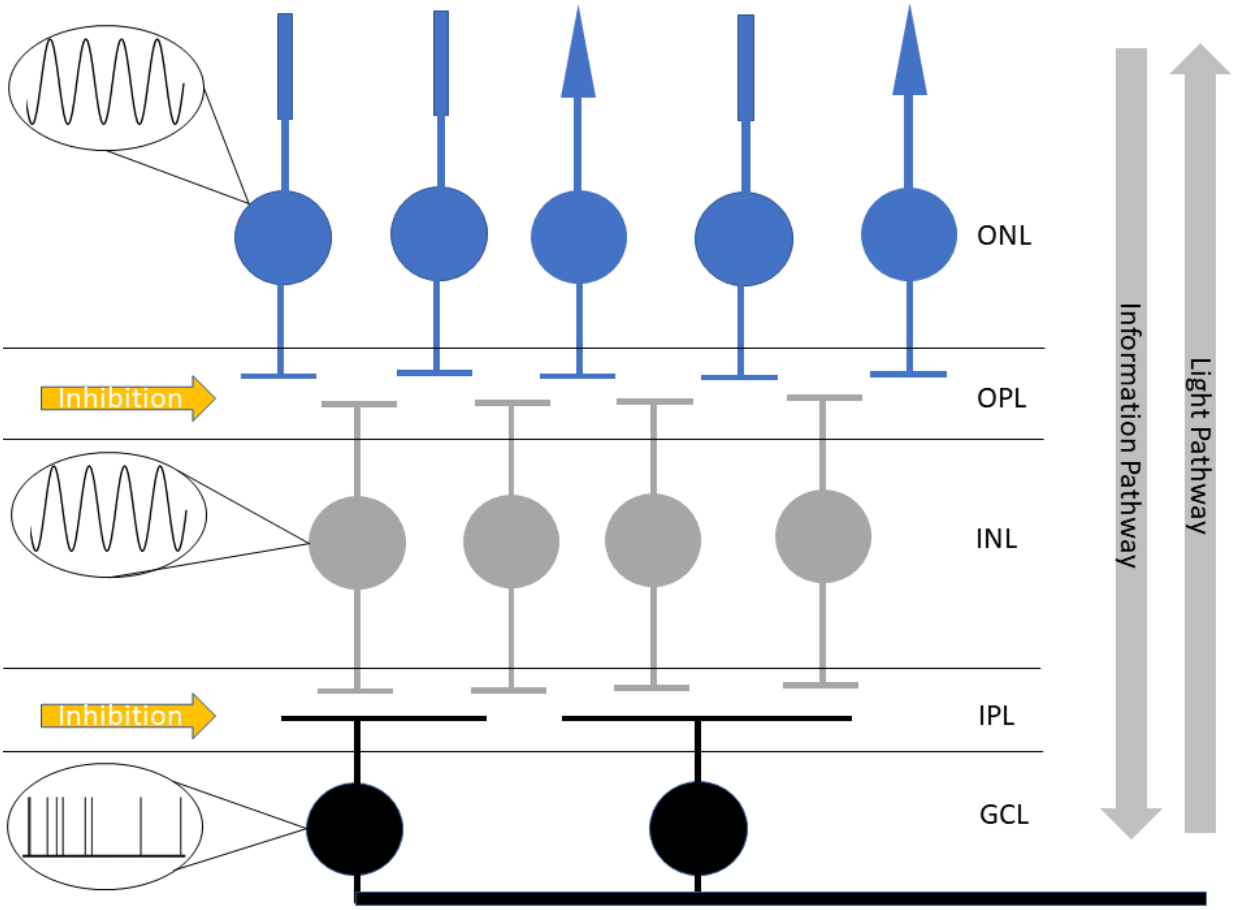

Figure 1.2: Basic excitatory retinal circuitry.

Light passes through the Ganglion and Inner layers before finally being detected by PRs (blue) in the outer nuclear layer. Information flow follows the reverse direction as light, from the PR through the outer plexiform layer (OPL) to the BCs (gray) of the Inner Nuclear Layer (INL). BCs then pass information through the inner plexiform layer (IPL) to the RGCs (black) of the Ganglion Cell layer, which then sends all visual information to the brain via the optic nerve. Note that before reaching the RGCs, all information is represented in an analogue fashion – it is not until the RGCs that information is represented purely by the digital sequences of binary spikes. Yellow arrows indicate inhibitory interneurons – Horizontal Cells of the OPL and Amacrine Cells of the IPL.

While the retina was used here as a specific example, this bottleneck is not uncommon in the early sensory systems, where information is locally processed before finally being

transmitted to the brain. Olfactory, auditory, and electroreception systems all have similar limiting features (limited by the nerves transmitting each signal to the brain, or the ratio of pre-cortical to cortical cells) (Kay and Sherman 2007). A first step in understanding neural processing in general, then, is to understand how these early sensory signals are processed before being sent to the brain. In doing so, one can note a few similarities in the cells of the early sensory system, and how they vary from higher order neurons. While the prototypical cortical neuron is a spiking neuron, representing information by the binary spiking and release of vesicles, many neurons in the early sensory systems operate by a continuous, graded membrane potential. Encoding information in this manner allows for increases in information transmission relative to spiking activity (Sengupta, Laughlin et al. 2014), at the cost of higher vesicular release rates (Laughlin, Howard et al. 1987). Returning to the retina, cells such as PRs and BCs operate largely in this fashion – it is not until the RGCs that information is purely represented in a binary spiking fashion. While there is still some uncertainty in how precisely this graded-to-digital switch occurs, one feature that plays a part in the release rates required to transmit graded information is the synaptic ribbon, a presynaptic organelle commonly found in early sensory cells such as PRs and BCs.

1.4: The Synaptic Ribbon and It's Properties

Also referred to as "dense body", the synaptic ribbon is an electron-dense protein complex tightly associated with vesicular release sites known as active zones (AZs). Morphological evidence suggests that the ribbon operates via tethering cytoplasmic vesicles and leading them to the AZ, although whether the conveyor belt analogy is entirely accurate remains open to debate (Parsons and Sterling 2003). Structurally, the ribbon is primarily composed of the RIBEYE protein which creates a scaffolding for vesicles to attach and is responsible for the electron density associated with imaging ribbons (Schmitz, Konigstorfer et al. 2000, Zenisek, Horst et al. 2004). Indicated in anchoring the ribbon to the plasma membrane and guiding functional AZ development are the accompanying proteins piccolo and bassoon, mutations in which are thought to cause free-floating ribbons that are no longer attached to the plasma membrane, as well as disturbances in vesicular release (tom Dieck, Altrock et al. 2005). Like conventional synapses, exocytosis from ribbon synapses is initiated by an influx of calcium. Voltage sensitive calcium channels will open in response to depolarization, allowing calcium to enter the cell and bind to calcium binding proteins that then activate a cascade of chemical machinery resulting in the fusion of vesicles with the plasma membrane and release of synaptic neurotransmitter (Ramakrishnan, Drescher et al. 2012). Notably, at the ribbon synapse, these

calcium binding proteins are situated directly adjacent to docked vesicles, aiding in rapid release and the temporal precision required of sensory systems (Gundelfinger, Reissner et al. 2015).

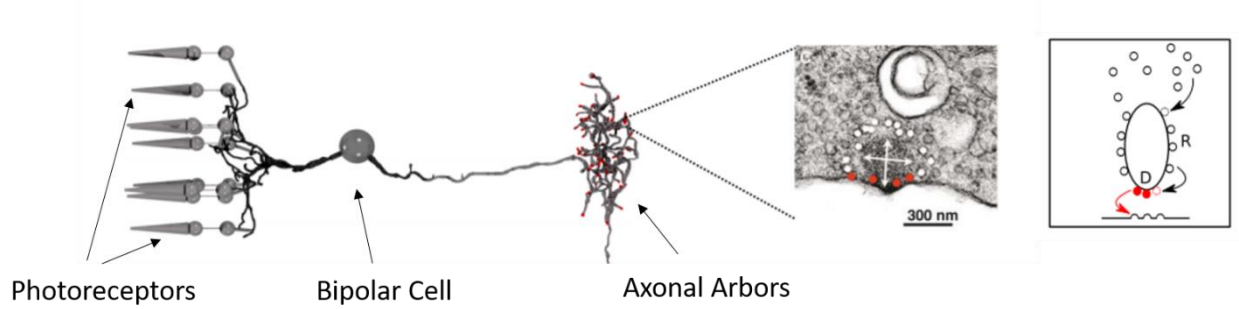

Figure 1.3: Reconstruction of a BC.

Blowup displaying an EM picture of the ribbon and associated vesicles (white circles are vesicles in the reserve pool, while red are docked vesicles of the readily releasable pool. Right) Cartoon of the ribbon synapses, showing the cytosolic, reserve, and readily releasable pools and the direction of vesicular movement. Taken from (Schroeder, James et al. 2019).

Morphological, electrochemical, and electrophysiological recordings of ribbon synapses have all indicated that the vesicles reside in three pools: the docked, or readily releasable pool (RRP), the reserve pool (RP), and free cytosolic pool (see **Fig1.2** for an example) (Von Gersdorff and Mathews 1994, Gomis, Burrone et al. 1999, Neves and Lagnado 1999, Parsons and Sterling 2003, Wittig and Parsons 2008, Lagnado and Schmitz 2015). The RRP consists of vesicles primed for release both spatially and functionally – that is, located directly adjacent to calcium receptors (von Gersdorff, Sakaba et al. 1998, Gomis, Burrone et al. 1999) and equipped with the biochemical machinery necessary for vesicle fusion (SNARES, synaptogmins, &c.) (Cho and von Gersdorff 2012). These vesicles can be rapidly released, exoctyosing all contents in as little as 200 ms (Von Gersdorff and Mathews 1994). The second pool is the reserve pool (RP). Larger than the RRP, the RP is thought to allow for fast – although not as fast as the release of the RRP – replenishment of the RRP. The last pool, the cytosolic pool, consists of free vesicles. Thus, docked vesicles are quickly released at the AZ, and vesicles from the RP function to refill the depleted pool, while the ribbon ‘grabs’ cytosolic vesicles to further refill the RP. While this strategy of ‘gather vesicles to the AZ’ is in no way unique to the early sensory system, the ribbon seems to amplify the strategies capabilities – not only can the rate of vesicle release reach the high rates required for transmitting graded signals,

it can maintain this release for extended periods of time with the aid of the ribbon (Baden, Euler et al. 2013).

High release rates are not the only property the ribbon provides, however. The nature of the vesicular pools in the ribbon – each pool composed of more and more vesicles with progressively slower movement rates – allow for another beneficial property for early sensory systems – adaptation (Burrone and Lagnado 2000, Baden, Euler et al. 2013, Johnston, Seibel et al. 2019). Ribbon synapses, stimulated with a constant drive, gradually decrease the rate of vesicular release – desensitizing to the stimulus. Rather than continually signaling the presence of a constant stimulus, then, the nervous system reduces metabolic cost by reducing the release rates in time. Notably, combining the desensitizing nature of the synaptic ribbon with inhibitory interneurons allows for the contrasting property – sensitization, wherein a cell responds to a constant stimulus with increasing release rates (Burrone and Lagnado 2000, Nikolaev, Leung et al. 2013). While this may seem at first counter-intuitive, this allows for efficient transmission of both stimulus increments and decrements. With the aid of the ribbon, then BCs can release excitatory glutamate onto RGCs in a way that signals not the raw stimulus itself, but the change in stimulus – an important aspect of compressing the visual information for transmission through the optic nerve.

A third property associated with ribbon synapses is multivesicular release (MVR), wherein multiple synaptic vesicles are released nearly simultaneously, with microsecond precision (Mennerick and Matthews 1996, Burrone and Lagnado 2000, Glowatzki and Fuchs 2002, Singer, Lassova et al. 2004, Neef, Khimich et al. 2007, Lagnado and Schmitz 2015). While MVR is not exclusive to cells of the early sensory system or graded cells – the phenomenon has been observed in the hippocampus (Tong and Jahr 1994), cerebellum (Auger, Kondo et al. 1998, Wadiche and Jahr 2001), and even somatosensory cortex (Huang, Bao et al. 2010) – it seems to be found in all cells containing the synaptic ribbon. Interestingly, ribbon synapses seem to show even more synchrony in vesicle releases, with MVR events occurring within microseconds. This type of MVR, with much tighter coordination between vesicles in release events, is aptly termed coordinated MVR (cMVR) (Singer, Lassova et al. 2004). Consequently, the ribbons itself has been suggested to elicit MVR. One theory is that the ribbon's tight clustering of vesicles near the AZ allows for compound fusion, wherein multiple synaptic vesicles fuse to one another before then fusing to the plasma membrane, effectively releasing all vesicles simultaneously. Regardless of the underlying mechanism, the functional purpose of MVR is little understood, for multiple reasons. Initial investigations of MVR have

relied upon either electrochemical or electrophysiological evidence, largely due to the high temporal resolution to distinguish true MVR events from non-MVR events offset in time. A consequence of these recording techniques is of course that the cell must be in direct contact with the recording apparatus – necessitating the use of *ex vivo* experiments. While a great deal of information can be obtained from this method, it prevents the understanding of how MVR operates in the context of the intact and *in vivo* circuit. How does MVR operate within the organism, and what functional roles can it play in information transmission?

1.5: Multivesicular Release

Ground-breaking work by Katz showed that neurons largely operate via a fast chemical signal in the form of synaptic neurotransmitter. 'Quanta' – discrete amounts of neurotransmitter reflecting the vesicular content of synaptic vesicles – are released from the presynaptic cell at the active zone (AZ) and activate receptors on the postsynaptic cell, resulting in the opening of ion channels (Del Castillo and Katz 1954, Heuser and Reese 1973). These channels function largely to depolarize or hyperpolarize the cell, and this membrane potential state then dictates whether vesicles are released from this cell to the next cell.

For a great deal of time, it was believed that the nature of this release was binary – a given synaptic release site can release a maximum of a single vesicle at once (Vere-Jones 1966, Zucker 1973). This notion corresponds well to the binary spiking activity exhibited in most neurons, which no doubt affected the 'uni-quantal' hypothesis' popularity. Taken all together, then, neurons were believed to transmit information to one another by modulating the rate of uni-quantal release (or analogously in the stochastic realm, the probability of a vesicle being released in a given time). Thus, vesicle release from N independent morphologically identified release sites and a probability of release p would obey binomial statistics – the number of vesicles released can be interpreted analogously to coin flipping – the number of heads out of N coin flips. The average number of vesicles released in this case would be Np, with corresponding variability Np(1-p). This hypothesis was strengthened by studies of in various brain areas, where experimentally observed release matches with binomial statistics (Korn, Mallet et al. 1982, Triller and Korn 1982, Dobrunz and Stevens 1997).

While elegantly simple in its formulation, and inviting many comparisons between brains and computers, this framework is not true. In fact, not only was it found that single release sites are capable of releasing vesicles nearly simultaneously (referred to as multivesicular release, or MVR), this phenomenon was observed in a multitude of areas of the nervous system, ranging

from the early sensory systems to higher brain areas such as somatosensory cortex and the hippocampus (Tong and Jahr 1994, Auger, Kondo et al. 1998, Wadiche and Jahr 2001, Christie and Jahr 2006, Higley, Soler-Llavina et al. 2009, Huang, Bao et al. 2010). This indicates that, rather than being an obscure exception to a fixed-vesicle release rule, MVR is a 'ubiquitous' process in the nervous system (Rudolph, Tsai et al. 2015). Developing more complex techniques with which to examine MVR in the nervous system is thus an important aspect of learning the significance of the phenomenon.

The initial observations of MVR in the nervous system were collected utilizing electrophysiological or electrochemical means. In these studies, the excitatory post-synaptic currents (EPSCs) of neurons receiving single inputs were measured. Decomposition of the resulting EPSCs showed that vesicles were being released nearly simultaneously – within the ranges of 10-100 μs. The tightly locked time in which these vesicles were released being referred to now as coordinated multivesicular release (cMVR), as it would seem the nearly simultaneous release of multiple vesicles reflects some sort of coordinated (and non-independent) release (Singer, Lassova et al. 2004).

1.6: Optical Examination of MVR

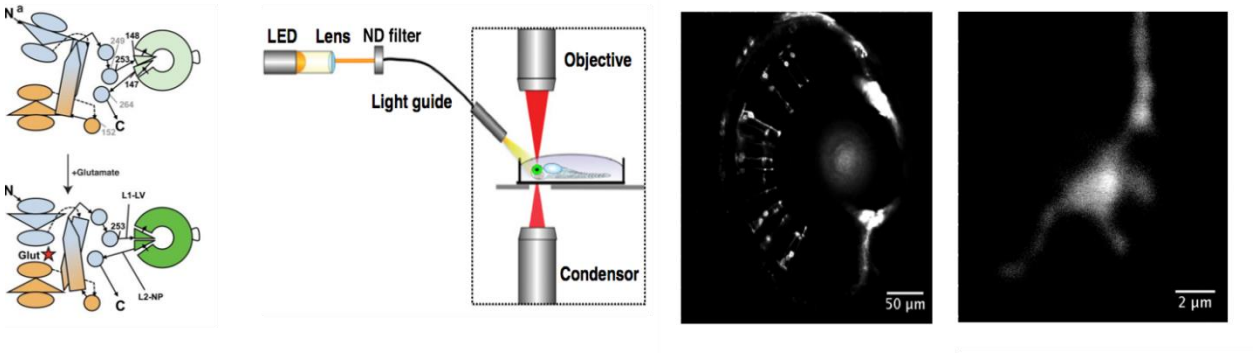

Figure 1.4: *In vivo* optical examination of glutamate release from zebrafish BC axon terminals.

Left) The iGluSnFR reporter, modified from (Marvin, Borghuis et al. 2013). The binding of glutamate induces a conformational change of the protein, allowing fluorescence. Middle) Recording setup for two-photon imaging. Full field stimuli are delivered through a LED light guide to the fish, embedded in agarose between a water objective and oil condenser. Right) Whole-field view of the zebrafish retina with BCs mosaically labeled. Far right) a zoomed in view of a single BC axon terminal.

Before we can begin to examine MVR in intact organisms, we must first construct a method which allows us to detect MVR *in vivo*. Two-photon imaging, while lacking the temporal resolution of the electrical recording techniques, has the advantage that it can be done *in vivo* in an intact organism and circuit. Rather than potentially damaging the underlying neural circuitry for electrophysiological recordings and confounding any general results regarding the function of the intact circuit, by using 2p imaging paired with genetic reporters, we can observe glutamate release from the intact circuit, or in the case of zebrafish – organism. This technique additional benefits from genetic targeting – specific cell types can be made to express the reporter (such as ribbon synapse containing BCs in our case), leaving other cell types unaffected. Additionally, an added benefit of the 2p technique over electrophysiology is that here we can actively image the synaptic release of glutamate, rather than surrogate measures used in electrophysiological techniques such as amperometry (Grabner and Zenisek 2013, White and Kim 2021), which require explantation for single-cell accuracy. While early fluorescent reporters lacked the kinetics and signal-to-noise ratio to accurately measure vesicular release, the recently introduced intensity-based glutamate sensing fluorescent reporter (iGluSnFR) (Marvin, Borghuis et al. 2013) has alleviated some of these issues. With maximum DF/F in the double digits and rise times of as little as 10 ms, iGluSnFR offers considerably more temporal resolution than earlier reporters. Combining this reporter with high speed, 1 kHz linescan imaging allows for enough precision to decompose fluorescence signals into units of individual vesicles, without requiring direct contact. Importantly, this allows for the interrogation of the phenomenon of MVR in a purely intact circuit in response to the systems natural stimulus.

1.7: Towards an Amplitude Code

Neurons have membrane time constant in the ranges of 1-50 ms. As such, the release of vesicles occurring in time windows of microseconds are likely to be considered as single events to the post-synaptic neuron, as voltage would not decay quickly enough between vesicle releases in MVR events to allow the cell to easily distinguish individual vesicles (REFS). MVR, then, violates the basic properties of rate coding – no longer are there simply two symbols (the presence or absence of an event), the synapse can release multiple vesicles simultaneously, expanding the range of possible responses from two to n+1, where n is the total number of possible quantal event types (i.e., uniquantal, 2-quantal, &c.), and the addition of one reflects the fact that the absence of an event is itself a symbol. Rather, MVR can be said to form an amplitude code, where information is contained not solely in the rate of a binary signal, but also the amplitude (corresponding to quantal content) of glutamatergic events. However, just

because ribbon synapses are *capable* of MVR does not necessarily mean that such an amplitude code is used by the nervous system. In order for this to be the case, the distribution of vesicles in an event must be dependent upon some aspect of the visual stimulus. While not likely, it is possible that MVR could be uninformative – if the distributions are not sufficiently shifted as a function of the visual environment. In order to test MVR's use in an amplitude code, then, we must first test the cell's ability to alter the distribution of event amplitudes dependent upon stimulus.

Neural systems contain inherent noise and stochasticity – ranging from the random isomerizations of PRs in the retina, Poisson noise in the spike times of neurons, to spike failures – when a spike from a presynaptic cell fails to induce a spike in the postsynaptic cell. All of these features could potentially be affected by the introduction of an alternative coding strategy. Could MVR be utilized informatively by the nervous system? How noisy is this amplitude signal? What potential advantages does this amplitude code convey over the traditional rate code? Could an amplitude code provide an additional useful mechanism in the BC-RGC graded-to-digital switch?

1.8: Aims

This work has three specific aims: to construct a method in which MVR events can be detected and quantified in *in vivo* settings; to show that, as opposed to the traditional rate-based hypothesis of neural information transmission, information can be transmitted between neurons by simply modulating the amplitude, or number of vesicles, in glutamatergic events; and to investigate how this information can be utilized by downstream neurons. In general, the hope is that this work can provide a better understanding of the vesicle code and how information transmission at the synapse can be understood, allowing for a further refinement of neural information processing techniques, especially with respect to the early sensory system and efficient coding.

Chapter 2: Methods

2.1. Zebrafish

In the recent years, zebrafish – *Danio rerio* – has become a commonly used animal model in various biological sciences. As a vertebrate, many aspects of its early sensory systems are shared across vertebrates, including mammals. Specifically, the zebrafish retina shows the same basic architecture as the mammalian retina, but with many additional benefits. Reproduction cycles of zebrafish are short, with a proportional maturation time. In the case of the retina, the large portion of the functional circuitry is fully formed at seven days post fertilization, when predatory behavior emerges (Easter and Nicola 1996). Zebrafish not only lay a large number of eggs and produce far more offspring in a quicker time than mammalian models, their larvae are also nearly transparent, an enormous benefit with respect to imaging. Combining this with zebrafish's easy genetic manipulation and introduction of cell-specific genetic reporters (Halpern, Rhee et al. 2008) makes them an ideal organism for the study of functional neuroscience.

2.1.1. Husbandry

Zebrafish were stored in a 14-10 hour light cycle. To aid imaging, fish were heterozygous or homozygous for the *casper* mutation, which results in hypopigmentation, and they were additionally treated with 1-phenyl-2-thiourea (200 μM final concentration; Sigma) from 10 h post-fertilization (h.p.f.) to further reduce pigmentation. All animal procedures were performed in accordance with the Animal Act 1986 and the UK Home Office guidelines, and with the approval of the University of Sussex Animal Welfare and Ethical Review Board, see (James, Darnet et al. 2019) for additional details.

2.1.2. Genetics

In order to record glutamatergic release from zebrafish BC terminals *in vivo*, we are using the intensity-based glutamate sensing fluorescence reporter iGluSnFR (Marvin, Borghuis et al. 2013). The reporter itself consists of a glutamate binding domain (using GltI taken from E. coli) implanted in a circularly permuted GFP. When glutamate binds to the reporter, it induces a conformational change in the cpGFP, increasing fluorescence. In order to make this reporter specific for Bipolar Cells (BCs) and Photoreceptors (PRs) in the retina, we place this transgene indirectly under the ribeye promotor – a promotor found in cells equipped with synaptic ribbons. Here, we are utilizing the GAL4-UAS system (Brand and Perrimon 1993, Scheer and Campos-

Ortega 1999). In brief, the GAL4 unit is under the control of a promotor (here the ribeye promotor), with 10 repetitions of the UAS enhancer embedded upstream to the iGluSnFR transgene. Activation of the ribeye promotor causes transcription of the GAL4 elements, which then bind to the UAS elements, causing the protein. Thus, only cells expressing the ribeye promotor will express iGluSnFR.

Initial reports of the kinetics of the reporter state rise times of tens of milliseconds, and a fall time of nearly 100 ms, with maximum in vitro dF/F of approximately 4.5 (Marvin, Borghuis et al. 2013). However, we found in zebrafish BCs the rise time to be as little as 1 ms, with a 60 ms fall time, and DF/F values of 3 – 5. It is likely that the expression of the reporter is then highly species and tissue specific. For example, measurements taken from the larval zebrafish lateral line system show significantly decreased signal-to-noise ratios, even when placed under the same promotor.

2.2. Two-photon Microscopy, Condenser

A fundamental aspect of the experimental work presented in this thesis utilizes fluorescence microscopy. In the most basic single-photon, or confocal, microscope, a laser with a specific wavelength is emitted onto a sample containing fluorescent dyes or reporters. Photons are then captured by the fluorophores, which then emit light of a different wavelength, which is then detected and transformed into an image (Yuste 2005). However, because a single photon is sufficient to activate the fluorophore, significant noise is added to the recordings due to activation above and below the sample, and the maximum depth of imaging is minimal. To counteract these deficits, two-photon microscopy was proposed (So, Dong et al. 2000, Svoboda and Yasuda 2006). Here, rather than the fluorophore absorbing a single short wavelength photon, it absorbs two longer wavelength photons in quick succession, again emitting light of a different wavelength. As the light emitted is of a longer wavelength and thus lower energy, this technique allows for imaging of tissues deeper than would be allowed with confocal microscopy. The other main advantage of the technique is its SNR. In 2p microscopy, the laser can be focused in such a way that only a precise portion of the sample receives enough photons to elicit fluorescence. In single-photon microscopy, the laser produces more background activation – as a single photon elicits fluorescence, any fluorescent molecules in the laser path are activated.

To further increase the SNR, we added an oil-based condenser below the sample, which greatly increases the amount of light collected. Rather than only collecting light from the

objective focused onto the sample, which as a result of light scattering can only detect a small portion of the emitted fluorescence, we collected light from below as well. This capturing of extra photons greatly increased the signal, allowing for the SNR required for the analyses presented in this work.

2.3. Light Stimulus

All light stimuli were generated using an amber light emitting diode (Thorlabs) with a peak wavelength of 590 nm and filtered through a 590/10 bandpass filter (Thorlabs). The light was delivered through a light guide placed directly adjacent to the eye of the fish and delivered as modulations around a mean intensity of ~320 nw/mm^2. Unless otherwise stated, all stimuli were delivered full-field at a temporal frequency of 5 Hz.

2.4. Data Analysis and Simulations

All imaging data excluding the immunohistochemical data were analyzed in Igor Pro (Wavemetrics) using custom-build code. The initial immunohistochemical analysis was done in DIANa in Fiji. Simulations of Poisson Processes and LIF models were performed in Python v3.2 on a desktop computer, and analysis of these data was performed in Igor Pro.

2.5 Information Theory

A variety of different information theoretical techniques are used in work. Here, I present a brief overview of the logic involved in the measurements utilized. Table 1.1 below lists the fundamental metrics of information theory, as well as their symbols, equations, and maxima.

Name	**Symbol**	**Equation**	**Maximum**
Information Content	$I(x)$	$-log[p(x)]$	∞
Entropy	$H(X)$	$E[I(X)]$	$log(N)$
Conditional Entropy	$H(X\|Y)$	$-\sum_{x\in X, y\in Y} p(x,y)log(\frac{p(x,y)}{p(y)})$	$H(X)$
Mutual Information	$I(X;Y)$	$H(X)-H(X\|Y)$	$min(H(X),H(Y))$
Specific Information	$I_2(X;y)$	$H(X)-H(X\|y)$	$H(X)$

Table 1: Common information theoretic metrics, their symbols, equations, and maxima.

In his groundbreaking 1948 paper, Shannon constructed a mathematical basis for which to quantify information transmitted between sources, laying the framework for information theory (Shannon 1948). Here, following along with Boltzmann's study of entropy in statistical mechanics (Gibbs 2015), he asserted that more information can be gained from less likely outcomes, introducing the metric known as information content or surprisal:

$$I(x) = -\log_b p(x)$$

Note the choice of base dictates the unit – the most commonly used base two yields bits. Logically following this nomenclature, base three is called 'trits', and base e and base ten yield nats and dats, respectively. Here x is an individual outcome from a discrete probability distribution X. The logarithmic function was chosen to agree with certain intuitive ideas about information: it is a nonnegative value, with the minimum of zero arising only when no uncertainty is involved; increasing probably of an event yields decreasing information, and the information contained in independent outcomes is additive. In fact, the logarithm is the only function that obeys all of these rules.

Information content then describes how much information one can gain from observing a single event based upon its probability. If, we instead chose to ask the question 'what is the average amount of information that can be gained after observing a symbol', we must average the information content across all observable symbols, leading to the definition of entropy:

$$H(X) = E[I(X)] = -\sum_{x \in X} p(x) log(p(x))$$

Where X is the probability distribution, x is an outcome from that distribution, and the E[ ] is the expectation operator. Note that here, while less probable outcomes contain higher information content than more probable outcomes, the less probable outcomes tend to have less of an effect on the system's entropy – their high information content is mitigated by their low probability. In fact, the maximum entropy distribution for a finite set of discrete symbols is a uniform distribution, where all symbols are equally probable. This results in a maximum entropy of $H(X) = log(N)$ – one bit for a binary uniform variable, where N are the number of distinct outcomes in the distribution.

While information content and entropy are viable methods to quantify how much information a random variable can transmit, they do not quantify how much one random variable can say about another random variable – a value quantified by mutual information. Before defining mutual information, however, it is useful to define conditional entropy, the amount of

information still required to determine the outcome of a random variable X after observing an instance of the random variable Y:

$$H(X|Y) = -\sum_{x\in X, y\in Y} p(x,y) log(\frac{p(x,y)}{p(y)})$$

Note that if X and Y are independent, then no information is gained about X from observing Y, and it thus still requires $H(X)$ bits of information to determine X. On the other hand, if Y completely determines X, then no additional information is required to determine X, and the conditional entropy is 0 bits. This then leads to a natural explanation of mutual information, the change in entropy of one variable after observing another:

$$I(X;Y) = I(Y;X) = H(X) - H(X|Y) = H(Y) - H(Y|X)$$

Note that mutual information is a symmetric quantity that can be defined numerous ways, but this is perhaps the most intuitive. For completely deterministic values, the mutual information is maximized to $min\{H(X), H(Y)\}$, while for independent variables the mutual information is 0. It is clear from the definition of mutual information that it represents the average amount of information observing one variable can tell about another. As such, it averages across all available symbols. In order to compute how much information the observation of a single symbol conveys about a separate random variable, we utilize the specific information (I2) (DeWeese and Meister 1999).

$$I2(X;y) = H(X) - H(X|y)$$

Like mutual information, specific information is a difference of entropies. However, unlike mutual information, specific information measures the change in uncertainty in one variable after having observed a specific symbol of a separate variable, rather than ANY symbol. Thus, mutual information is a vector rather than a single value, with one element for each supported symbol. Another defining difference between the two metrics revolves around the probability of observing each symbol type. While mutual information weights the information conveyed by a specific symbol by its probability, specific information does not – it simply indicates the amount of information a symbol contains about a separate random variable regardless of the probability of actually observing that symbol. For example, the appearance of a zero in a binary system may convey a large amount of information about a second random variable, thus yielding a high specific information. However, this symbol may be very unlikely to occur, meaning that it only slightly affects the mutual information. Linking the two metrics even more, one manner in which

to compute the mutual information of two variables is to take the inner product of the specific information vector with the probability of observing each sequence in that vector:

$$I(X;Y) =< I2, pR >$$

Where I2 is the specific information vector and pR is the vector representing the probability of observing each symbol, and <> is the inner product.

Note that this is not the only definition of specific information. The alternative metric for mutual information, also referred to as I1 and given by the equation:

$$I_1(X;y_j) = \sum_{x \in X} p(x_i|y_j) log[\frac{p(x_i|y_j)}{p(x_i)}]$$

While this formulation seems to be more common in the neuroscience field (Eckhorn and Popel 1975, Theunissen and Miller 1991, DeWeese and Meister 1999), I have opted to use the former definition due to its ease of interpretability. While both metrics attempt to quantify the information gained by a single symbol, I2 has the added benefits of both being additive and equating simply to a difference of entropies, as seen in the previous equation.

2.5.1. Implementing information theory

We took the standard approach to estimating information theoretical metrics from event-based neural data (Strong, de Ruyter van Steveninck et al. 1998). Following the identification and quantization of events, I discretized the response by creating time bins, where the length of each bin (between 10 – 20 ms) is set in such a way that no two events can occur in any single bin. In doing so, we construct a multinary sequence of events, where bins with no event are given the value a zero, and bins with events are given the value corresponding to each event's quantal content. For example, the five bins proceeding a stimulus might give the sequence: **05030**, where the first, third, and fifth bins contained no events, the second a five-quantal event, and the fourth a three-quantal event.

Note that information theoretic metrics are notoriously difficult to estimate, and the data we are analyzing is particularly susceptible to bias. In the traditional binary analysis, a five bin response could result in 2^5 different responses, while a multinary signal with up to ten symbols could result in a potential 10^5 different responses. For this reason, the experimental work did not attempt any analysis of the sequence of synaptic events – we are simply estimating the conditional entropy of each observed symbol, independent of the symbols that occur previously

or the time at which these symbols occurred. While each AZ produces a sequence of events, we are not interested in their order, only the distribution of single events within this response. As such, the number of potential responses is significantly reduced, as is proportionally the bias. Here, I mostly opted for the traditional empirical estimates of probability distributions, where the probability of any event is based almost entirely on the observed frequency. However, two additional manner were constructed to prevent bias. The first was the addition of a small Jeffrey's prior (Gelman, Carlin et al. 2013), equivalent to placing half an 'observation' of each symbol in each bin. The other way we reduced bias was restricting the analysis to events with enough observations to reliably estimate their frequency, ensuring that the number of responses collected was greater than or equal to the number of distinct responses (Panzeri and Treves 1996, Panzeri, Senatore et al. 2007). For instance, the number of distinct responses for the specific information analysis was under twelve, corresponding to events consisting of from zero to 11 quanta, while a total of hundreds of responses were collected. Consequently, many events that did not occur frequently enough were discarded from the analysis. While more advanced techniques exist for information theoretical inference, these often time place a strong emphasis on prior probabilities for smaller data sets, ironically potentially increasing bias, and they were not adopted.

2.6. Poisson Processes

While the focus of this work is not on Poisson Processes, they are utilized to an extent that a rough description of the process is necessary. A Poisson Process is a type of stochastic point process – a mathematical formulation of the random spread of points on some measurable space (Kingman 1992, Resnick 1992). While multiple definitions of a Poisson Process exist (and with varying levels of mathematical complexity) the simplest and most intuitive interpretation of a PP is as a counting process. For the sake of simplicity, we will restrict our focus to Poisson Processes in time on the positive half real line – such as the time at which neural action potentials occur after stimulus onset at $t = 0$. If we assign some measurable function $N(t)$ that 'counts' the number of points that have occurred in the window $(0, t]$, and this measure has the following properties:

1) $N(0) = 0$; the count starts at zero
2) The distribution of events occurring in non-overlapping windows are independent
3) $N(t) \sim Poiss(\lambda t)$; the probability of k events in a window of length t is $p(k) = \frac{e^{-(\lambda t)}(\lambda t)^k}{k!}$

Then the process, denoted $\{N(t): t \geq 0\}$, is a Homogeneous Poisson Process with parameter λ. Note that PPs can also be defined as a point process by their independent exponentially distributed inter-arrival times, but either definition implies the other (See appendix for details).

The above definition considered the time-invariant case, where the rate of events is constant in time. This is not a necessity, as we can see with the Nonhomogeneous Poisson Process. Here, the rate of events varies in time according to an instantaneous intensity function $\lambda(t)$. In this case, we find that the number of events in a time window $(0, t]$ is distributed according to a Poisson distribution with parameter $\Lambda(\mathrm{t})$, where

$$\Lambda(\mathrm{t}) = \int_0^t \lambda(s)\, ds$$

Note that the Homogeneous case is just a subset of the Nonhomogeneous case, with $\lambda(t) = \lambda$, and the NHPP can be interpreted as a rescaled HPP equipped with an 'intrinsic age' (Cinlar 2013).

While PPs have been successfully used to model various neural phenomena, we note that the basic forms of PPs are not well suited to modelling MVR. As a consequence of the definition of PPs, no two single events may occur simultaneously (this can be understood by the fact that $\lim_{h \to 0} P(N(h) > 1) = o(h)$, with little o notation indicating that the probability approaches zero faster than the value h. Thus, the simultaneous release of multiple vesicles in MVR violates one of the most basic properties of PPs, and an extension of the ideas is necessary to effectively model the phenomenon.

2.6.1. Splitting Poisson Processes

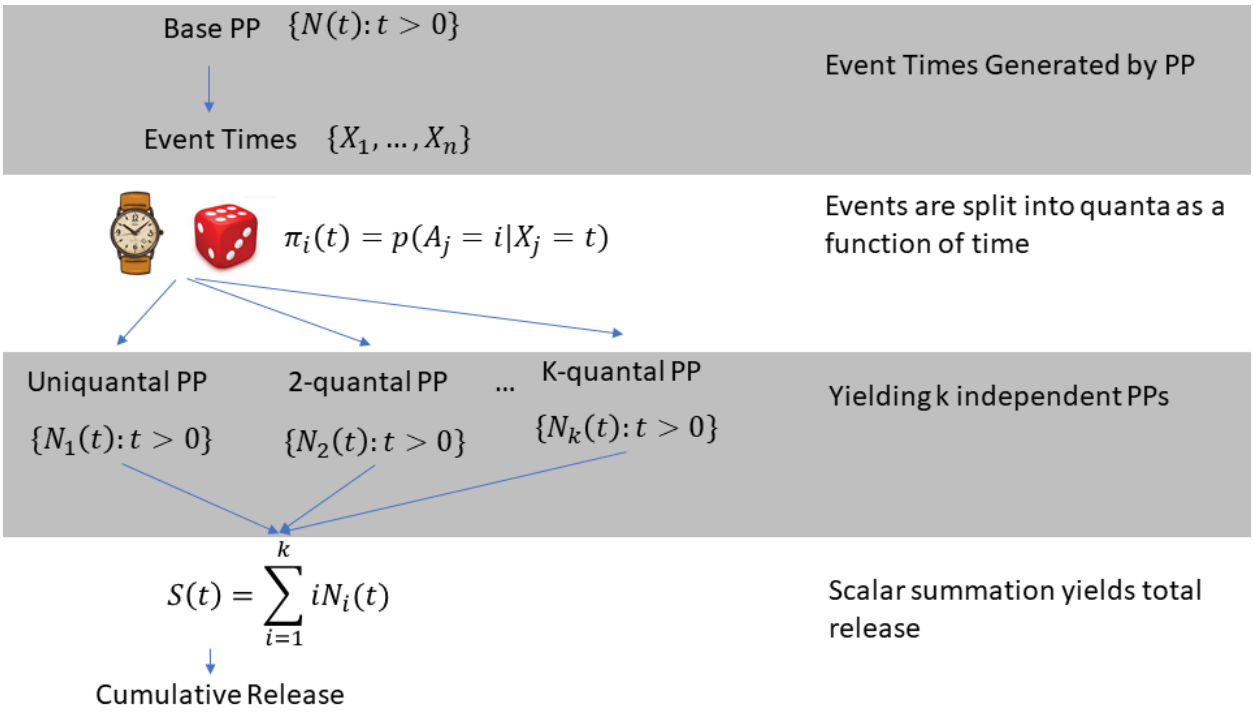

Figure 2.1: Graphical Representation of Poisson Splitting.

A base PP is split into k sub-processes, all of which are independent and Poisson, allowing for analytical computation of the moments of cumulative release.

Informally put, Poisson Splitting is a property of PPs that states that a base PP can be stochastically split into one of k subtypes with a possibly time-dependent probability, and the resulting subprocesses will all be independent of one another and Poisson. By assigning a quantitative value to each type corresponding to quantal content, this allows us to introduce time-dependence to mimic MVR, as well as maintain a simple solvable equation for the moments of cumulative vesicle release.

Mathematically, Poisson Splitting states the following: given some base Poisson Process $\{N_B(t): t > 0\}$ with instantaneous intensity function $\lambda(t)$, we can stochastically split each event into one of k mutually exclusive and exhaustive subtypes, given by $\{N_i(t): t > 0\}$, with some splitting probability $\pi_i(t)$. The resulting subprocesses are independent from one another, each with intensity function:

$$\Lambda_i(t) = \int_0^t \lambda(s)\, \pi_i(s)\, ds$$

Because we can assign an integer value to each subtype (uniquantal events, two-quanta, etc.), the cumulative release can simply be described (in a somewhat bastardized notation) by

$$S(t) = 1\Lambda_1(t) + 2\Lambda_2(t) + \ldots + k\Lambda_k(t) = \sum_{i=1}^{k} i\, \Lambda_i(t)$$

and the expected cumulative release as

$$\begin{aligned} E[S(t)] &= E\left[\textstyle\sum_{i=1}^{k} i\, \Lambda_i(t)\right] && \text{by definition} \\ &= \textstyle\sum_{i=1}^{k} E[i\, \Lambda_i(t)] && \text{by independence} \\ &= \textstyle\sum_{i=1}^{k} iE[\, \Lambda_i(t)] && \text{by linearity of expectation} \\ &= \textstyle\sum_{i=1}^{k} i\Lambda_i(t) \end{aligned}$$

The variance can likewise be defined as $Var[S(t)] = \sum_{i=1}^{k} i^2\Lambda_i(t)$. Note that simulating these processes can be accomplished in two ways – by first computing the intensity functions for each sub-process and then sampling directly from those NHPPs, or by first sampling event times and then stochastically splitting based upon their time.

2.6.2. Simulating Poisson Processes

Homogeneous PPs are simulated simply by drawing interevent times from an exponential distribution with parameter λ^{-1} based on the definition of a Poisson Process. In order to generate event times from NHPPs, we utilized Çinlar's Method (Cinlar 2013). Here, the dependent variable can be transformed to a standardized unit according to the intensity function $\Lambda(t)$, and event times are then generated according to the following algorithm:

1) Set $s = 0, tTot = 0$
2) Generate $u \sim U(0,1)$
3) $s -= ln(u)$
4) $t = inf\{v: \Lambda(v) \geq s\}$
5) Deliver t
6) $tTot += t$
7) Repeat (2-6) until $tTot > T$

Here, $\Lambda(v)$ is the intensity function of the PP, inf is the infimum – the largest values of v less than or equal to the value s, and T is the total length of the window being simulated.

2.7. The Leaky Integrate and Fire Model

In order to simulate spiking activity of cells postsynaptic to ribbon synapses, I utilized a Leaky Integrate-and-Fire (LIF) Model of neural activity (Tuckwell 1988, Burkitt 2006, Burkitt 2006) receiving stochastic input. In this model, voltage is governed by the differential equation:

$$C_m \frac{dv}{dt} = \big(v_{rest} - v(t)\big) + \big(v_E - v(t)\big) g_E S_{in}(t),$$

Where C_m is capacitance, g_E is conductance, V_E is the excitatory reversal potential, and v_{Rest} is the resting membrane potential. The synaptic input, S_{in}, is defined by the convolution of an alpha current impulse function with the stochastic Poisson input s(t),

$$S_{in}(t) = \alpha \ * \ \mathrm{s(t)} = \sum_{i=1}^{n} \alpha(t - t_i)$$

Where s(t) is the sequence of Dirac delta function describing the timing of each event in a realization of the Process, $\alpha(t) = exp\left(\frac{-t}{\tau_{decay}}\right) - exp\left(\frac{-t}{\tau_{rise}}\right)$ describes the time course of synaptic current, and t_i are the times of the occurrences of the Poisson Process.

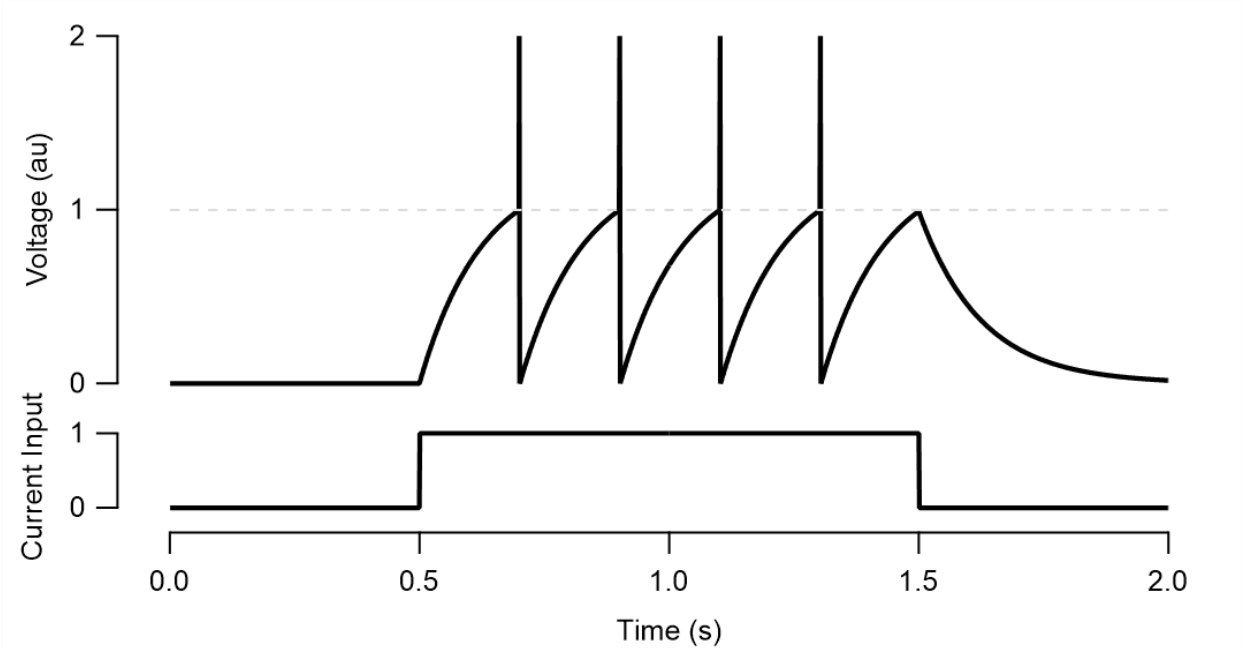

Figure 2.2: Example of the LIF Model.

A cell receives a constant current input (bottom). This raises the cells voltage (top) until a spiking threshold is reached (gray dashed line). At this point, the cell spikes (indicated by the voltage pulses) and is reset back to resting potential. Note that the more depolarized the cell is, the less effective the current is at increasing the cell voltage. In the absence of input, the cell decays naturally back to its resting potential, as shown by the failure to reach spiking threshold for the last event.

Once the membrane potential reaches a threshold θ, the cell fires and the voltage is immediately reset to the reset potential v_{Reset}. As is common in using the LIF model, we transformed the formula by introducing the membrane time constant $\tau = RC$, with values ranging from 1 to 50 ms. To aid in mathematical simplicity, voltage values were additionally transformed from their scientific units of volts into arbitrary units ranging from zero to one, and all remaining parameters were correspondingly altered. For instance, in the traditional model, where $v_{Rest} = -60\,mV$ and $\theta = -40$ mV, the excitatory reversal potential is correspondingly shifted from values around $0\,mV$ to approximately 3.75. As an important aspect of my utilization of the LIF model requires precisely defining the number of simultaneously released vesicles (or total released vesicles in a perfect integrator) required to generate a spike, we altered the excitatory conductance in order to adjust the number of vesicles required to generate a spike while holding constant the membrane time constant.

Chapter 3: A method for *in vivo* counting of vesicles in MVR Events

3.1. Introduction

Despite the apparent advantages electrophysiological or electro-chemical recordings display over imaging, most electrophysiological or electro-chemical data cannot be measured *in vivo* in fully intact circuits, generally requiring *ex vivo* or cultured cells such that a physical contact can be made between the cell and the recording electrode. Even traditional electrophysiological recordings of the retina generally require explants, which tend to tear the tissue and alter functional connectivity. Other methods used to quantify vesicle release, such as amperometry, are notoriously difficult to perform in smaller cells, with experiments often done on particular large neuron such as in the Calyx of Held or Rod Bipolar Cells of the goldfish retina (Neves and Lagnado 1999, Grabner and Zenisek 2013). Thus, while MVR has been observed in distinct brain areas and systems, it has been difficult to attribute a functional significance to the process. How does MVR operate in the intact organism? What advantages may the phenomenon impart upon biological processes? In order to answer these questions, we must first develop a strategy to observe and quantify MVR in whole organisms. To do this, we turn to imaging of genetically encoded indicators (Lin and Schnitzer 2016), allowing for both genetic targeting of cell types, as well as *in vivo* recordings.

While imaging using two-photon microscopy allows for the advantage that it can be undertaken in intact neural circuits – or even intact, awake, and behaving organisms – it nonetheless suffers from its own drawbacks. An appropriate fluorescent reporter must be engineered and constructed, the spatio-temporal resolution is limited by mechanical aspects, and signal-to-noise ratios (SNR) can be low. Recent advances have mitigated these drawbacks, however. Here, I am utilizing the intensity-based glutamate sensing fluorescent reporter iGluSnFR (Marvin, Borghuis et al. 2013) to image glutamate release from zebrafish Bipolar Cell (BC) axon. The reporter itself consists of a circularly permuted GFP inserted into the glutamate binding domain GtlI from E. Coli. Thus, glutamate binding causes a conformational change in the GFP, resulting in an increase in fluorescence. The result is signal-to-noise ratios of up to ten $\Delta F/F_0$ or above - improved even further by the use of a condenser to collect more light - as well as fast kinetics, with rise times of as little as 1 ms in zebrafish BC axon terminals. To increase temporal resolution, I sacrificed spatial resolution, allowing for recording of single release sites at 1 kHz temporal resolution. These optimizations have allowed for the resolution required to

count vesicles within glutamate release events, analogous to what has been done in electrophysiology.

Here, I developed a method allowing for decomposition of glutamatergic events into units of individual vesicles from imaging data. Using the iGluSnFR reporter bound to the ribeye promotor expressed in zebrafish Bipolar Cells (BCs) and recording high-speed 2p linescans allows for the counting of vesicles within glutamatergic events. Using both experimental and simulated recordings, I verify the method and show its advantages and limitations. While incapable of reaching the temporal precision found in electrophysiological or electrochemical recordings, I nonetheless am able to distinguish individual glutamatergic events separated by as little as 10 ms. The described decomposition is sufficiently precise to estimate not only the vesicular content of glutamatergic events, but the time at which these events occurred. Thus, the method outputs the traditional vector **E** of event times (similar to the vector of spike times), but also the vector **AQ**, the estimated number of vesicles composing the event. In the following chapters we will see how this representation is advantageous for scientific analysis. This analysis allows for quantal decomposition of vesicle release in intact organisms using iGluSnFR, a technique previously only available using electrophysiological recordings.

3.2. Methods

3.2.1. Method Overview

I created an analysis package in Igor Pro (Wavementrics) to detect and quantize glutamatergic events from two-photon linescans across the terminals of bipolar cells expressing iGluSnFR (see **Fig3.1** for an example of the zebrafish retina expressing iGluSnFR in BCs). The data takes as input a line scan matrix – an ***x*** x ***L*** matrix in which ***x*** is the number of pixels per line and ***L*** is the number of lines. All data simulated and analyzed here were carried out with a temporal frequency of 1 kHz – recording each of the 128 pixels in the linescan 1,000 times per second. The analysis defines regions of interest (ROIs) corresponding to active zones, detects and measure glutamatergic events and provides a final output of a set of ***E*** event times and **AQ** estimated quanta for each event in each active zone. The analysis is composed of six main steps:

1. ROI detection by spatial decomposition

2. Time series extraction by weighted averaging
3. Baseline correction and calculation of ΔF/F
4. Identification of events by Wiener deconvolution
5. Extraction of events
6. Amplitude clustering and quantal time series

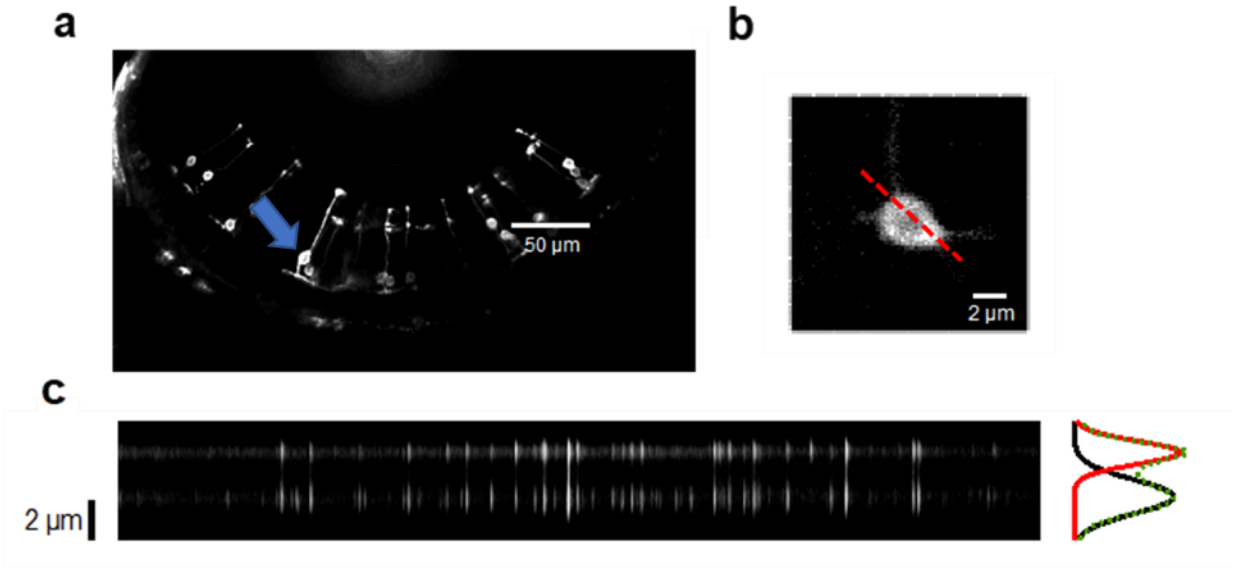

Figure 3.1: 2p Imaging of zebrafish BC glutamate release.

a) View of the zebrafish retina showing mosaic expression of the iGluSnFR transgene. Blue arrow indicates a single BC, pointing to the cell body. **b)** Zoom in of a bipolar cell terminal, with the recording line shown in red. **c)** Linescan matrix of the terminal in b), showing two active zones. The right portion shows the temporal average of the trace.

An overview of the main steps involved in the analysis is shown in **Fig3.2.**

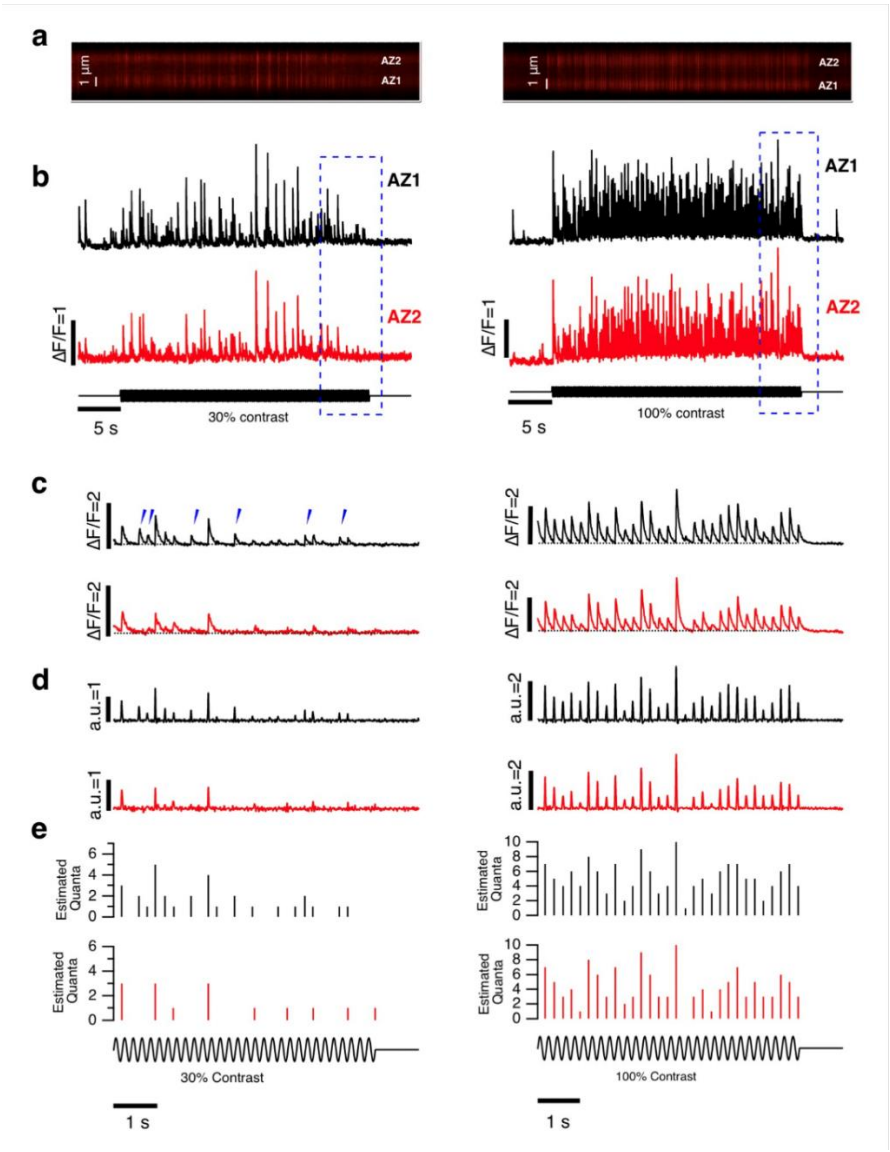

Figure 3.2: Overview of major steps in analysis.

a) A linescan showing changes in iGluSnFR fluorescence in a synaptic terminal in response to a full-field stimulus modulated at 5 Hz (sine wave). The left-hand side shows the profile of two active zones of a terminal stimulated at 30% contrast and the right-hand side the profile of the same terminal stimulated at 100% contrast. **b)** Traces extracted from the linescans shown in a. **c)** Expansion of the period shown boxed in b. Note that at 30% contrast there is events at one active zone that do not coincide with events at the other (highlighted by arrowheads), whereas 100% contrast drives responses in both active zones reliably over all cycles of the 5 Hz stimulus. Analysis shown in d and e were extracted from the boxed area. **d)** Deconvolved trace using the estimated filter (Wiener). **e)** Estimation of the number of quanta per event. The stimulation protocol is represented below.

3.2.2. Roi Detection

To define ROIs within line scans, I first noted Fick's second law of diffusion describing the spatio-temporal diffusion of a substance:

$$\frac{\partial C}{\partial t} = D\frac{\partial^2 C}{\partial x^2}, \tag{1}$$

where the change in concentration (***C***) can be described in one dimension (***x***) over time (***t***) with a diffusion coefficient (D). If *N* is the initial number of glutamate molecules released instantaneously at a point (corresponding to a synaptic release site), a solution can be in the form:

$$C(x,t) = \frac{N}{\sqrt{2\pi\sigma^2(t)}} * \exp\left[-\frac{x^2}{2\sigma^2(t)}\right], \tag{2}$$

where

$$\sigma^2(t) = 2Dt. \tag{3}$$

This function is recognizable as a normal distribution with variance a monotonically increasing function of time. Here, lines were sampled at intervals of 1 ms, making it difficult to distinguish any spatio-temporal dynamics. Because of this, as well as the Gaussian form of the microscopes point spread function, I therefore measured the spatial profile as a temporal average of the fluorescence signal along the linescan and fit this average to a sum of Gaussians, where each Gaussian component can be considered its own point source corresponding to an active zone. One of these fluorescence profiles is shown in **Fig3.3** defining two nearby active zones. Using a GUI, the user specifies the peak(s) in the profile by placing one or more cursor s and I then use IgorPro's built-in curve fitting routines to fit the temporal average to the function k(x):

$$k(x) = \sum_{i=1}^{n} \frac{A_i}{\sqrt{(2\pi\sigma_i^2)}} * \exp\left[\frac{-(x-\mu_i)^2}{2\sigma_i^2}\right] \quad (4)$$

under the constraints

$$\mu_i \in [c_i - \delta, c_i + \delta],$$
$$A_i > 0,$$

where *c* is the set of cursor locations, *μ* the set of component means, *A* the set of component amplitudes, σ^2 is the set of variances, $\boldsymbol{\delta}$ is a small value allowing for errors in user cursor placement, and *n* is the number of placed cursors. Here, cursors were placed to allow for the fitting of the variables *μ, A,* σ^2. The potential problem of overfitting the intensity profile with multiple Gaussians was avoided by restricting the number of components to the number of placed cursors. This process could be repeated with different initial estimates of the locations of the peaks (i.e., the number of cursors and their positions) until the error function reaches a threshold or until the fit is acceptable. User input at this stage allowed us to limit the analysis to active zones with distinct peaks and FWHM <1.5 μm, thereby reducing the possibility of conflating signals from multiple active zones.

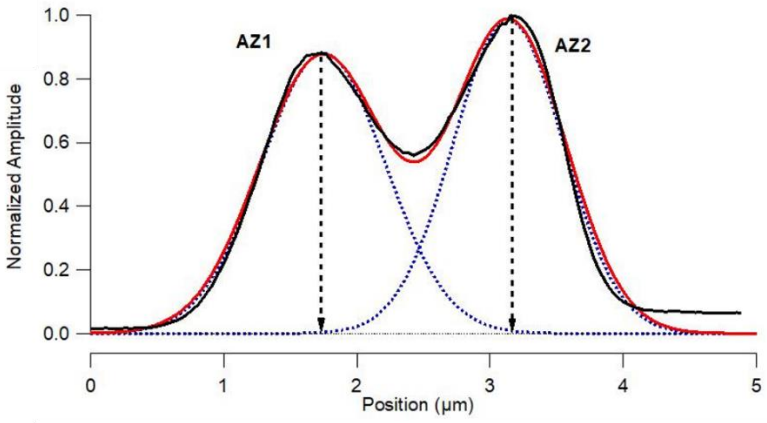

Figure 3.3: Spatial demixing of iGluSnFR signals from neighboring active zones.

The temporal average (black trace) is fit with the sum of two Gaussians (red trace). Dashed blue line shows individual components. The FWHM values were 1.1 mm (left) and 0.96 mm (right).

3.2.3. Time Series Extraction

Once each spatial component had been defined, a time-series for that component, F(t), was computed as the weighted average of the raw linescan matrix with the spatial filter estimated in step 1:

$$F(t) = \sum_{x} F(x,t)k(x)\,, \tag{5}$$

where F(x,t) is the raw linescan matrix and k(x) is the Gaussian component for the ROI, and the summation is done numerically across all 128 pixels of a linescan. Increasing the weight of pixels located towards the center of the spatial profile (**Fig3.3**) allows for significant denoising. Thus, the signal extracted from pixels nearer the center of the spatial profile have higher amplitudes, and thus higher signal-to-noise, than those extracted from the edges of the profile. While utilizing Gaussian decomposition in such a manner does have the potential to contaminate an ROIs data by including activity from neighboring ROIs, I found that the technique provided better SNRs than manually selecting bounds for each ROI to average over uniformly. For the vast majority of recordings consisting of multiple ROIs, the distance between ROIs was enough to adequately disentangle the signals using the above method.

3.2.4. Baseline Correction and Calculation of $\Delta F/F_0$

Bleaching of iGluSnFR sometimes occurred during an observation episode and was usually corrected using a linear function of time F(t). The iGluSnFR signal used for all analysis was the relative change in fluorescence, $\Delta F/F_0$, calculated from the bleach-corrected signals. The most frequent value (i.e., the baseline) of the trace was used as F_0.

3.2.5. Identification of Events

Release events within an active zone were identified by their characteristic kinetics using a Wiener filter. The time-course of 101 iGluSnFR transients that were clearly separated in time from other events are shown in **Fig3.4** and could be described by a function of the form:

$$h(t) = A * exp\left[-\frac{t}{\tau_f}\right] * \left(1 - exp\left[-\frac{t}{\tau_r}\right]\right), \tag{6}$$

where the event is taken to occur at time zero and where *A* describes the amplitude of the event and τ_r and τ_f and are the time constants for rise and fall in the signal, respectively. I found that transients at most synapses could be accurately described using a kernel with parameters of τ_f = 0.06 s and τ_r = 0.001 s. These parameters were relatively invariant for transients of different amplitudes (**Fig3.4**), indicating that the reporter operated linearly over the range of glutamate concentrations that I observed. These observations strongly suggest that these signals reflect a linear time-invariant system (LTI), fulfilling the assumptions required for the use of Wiener deconvolution. The filter described by equation 6 therefore allowed me to both "denoise" signals and estimate the underlying input. The result of the Wiener deconvolution was a time series in which glutamate release events were described approximately as Dirac-δ impulse functions of varying amplitudes, as shown by traces in **Fig3.2d.**

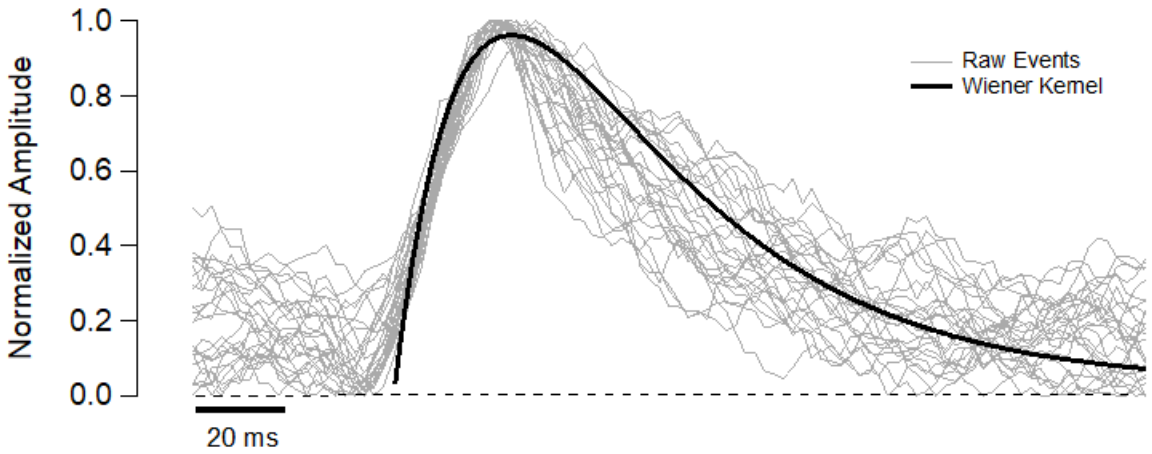

Figure 3.4: **The Wiener kernel used for deconvolution.**

Representative example of raw events overlayed with a Wiener kernel given by equation 6. Note that the time decay constant of the fluorescence signal is approximately 50 ms. A total number of 101 raw events were averaged for estimating the Wiener kernel filter.

3.2.6. Extraction of Events

Although the use of Wiener deconvolution significantly improved the signal-to-noise ratio, it was still necessary to set a threshold to distinguish events from noise. A second example analysis highlighting this key step is provided in **Figs3.5** and **3.6**. Fig **3.5a** shows a kymograph of a

linescan through a terminal in which there were two sources of glutamate, (active zones 1 in red and active zone 2 in black), together with the activity time-series for each obtained after spatial demixing (step 1). The corresponding traces after Wiener deconvolution (step 1) are shown in **Fig3.5b,** where it can again be seen that glutamatergic events varied widely in amplitude. The baseline in the deconvolved traces was not, however, noiseless, making it necessary to set a threshold for counting a deviation in this signal as an event. To choose this threshold, I first examined the distribution of values in the deconvolved trace. Across all experiments, these distributions were consistently Gaussians centered at or very close to zero, except for a small tail of positive values. I therefore used the standard deviation of a Gaussian fit to the distribution to set the threshold at which positive values in the deconvolved trace were considered to be significant (i.e., to reflect iGluSnFR events). The thresholds I used were 3-4 standard deviations above the baseline, as shown by the dashed blue lines in **Fig3.5b**. Events were then timed at the local maximum in the deconvolved trace above this threshold, as shown by the dashed vertical lines in the expanded traces in **Fig3.6**. The activity within an active zone could then be described by a vector **E** of event times and **A** of event amplitudes.

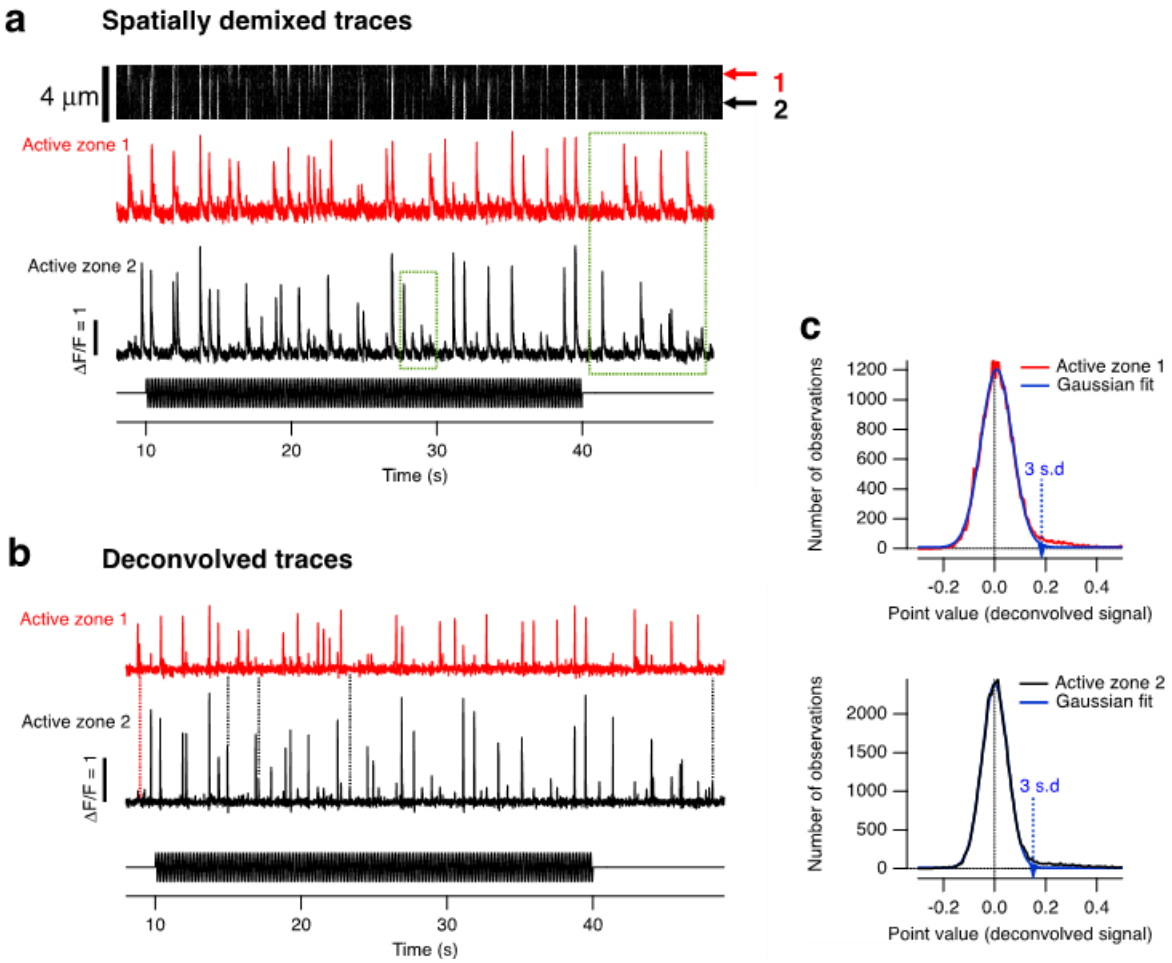

Figure 3.5: Differentiating events from noise.

a) The kymograph (top) shows the intensity profile along a line through a single terminal. The red and black traces (middle) show the time-course of the iGluSnFR signal over the two active zones marked to the right of the kymograph and the stimulus is shown immediately below. After 2 s there was a switch from constant illumination to full field modulation at 20% contrast, 5 Hz. The signals were demixed using a weighted sum of two Gaussians fit to the intensity profile along the line-scan, according to steps 1.1 and 1.2. Two sections of the record (green boxes) are expanded in Fig3.6. Note variations in the amplitude of glutamate transients. **b)** The results of Weiner deconvolution applied to the traces in **a** using the kernel shown in Fig3.3. The dashed red line shows an event in active zone 1 that did not coincide with an event in active zone 2, and the dashed black lines highlight events in active zone 2. **c)** The distribution of values in the traces shown in **b** (active zone 1 in red to the top and active zone 2 to the right in black) together with a fitted Gaussian (blue). The threshold of 3 sd above the baseline is indicated by the dashed blue arrow.

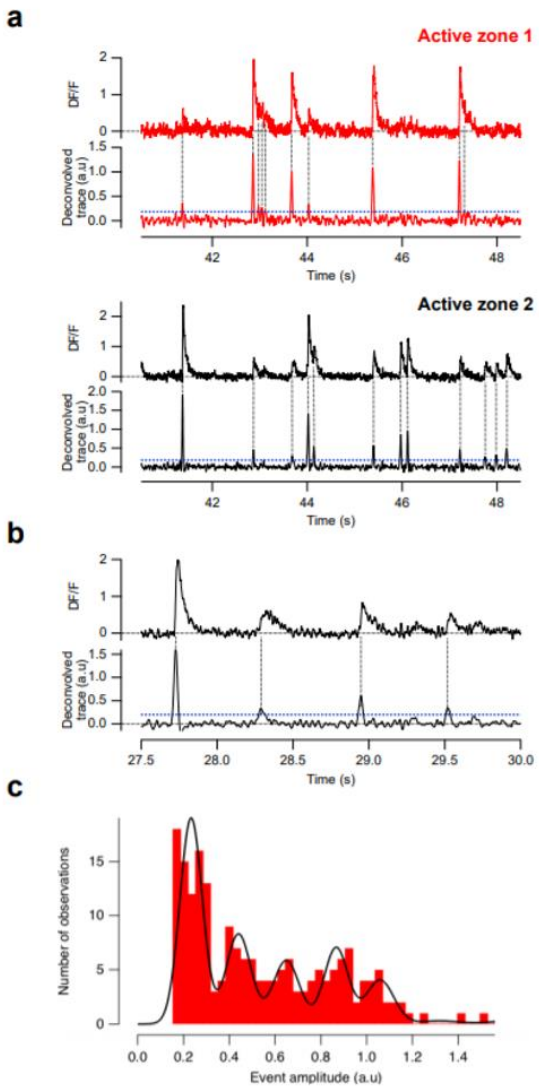

Figure 3.6: Examples of events detected.

a) Activity in active zone 1 (red) and active zone 2 (black), with deconvolved traces immediately below the iGluSnFR signals. The period of activity corresponds to the large green box in Fig3.5a. Thresholds for differentiating events from noise are shown by dashed blue lines, and the dashed vertical lines link the event in the deconvolved trace to the iGluSnFR trace. Note the burst of one large and then three smaller events around 43 s in active zone 1. Note also the deviation in the iGluSnFR signal from active zone 2 at 43 s which was not counted as event because it was small and slowly rising. **b)** Activity in active zone 2 over the period shown by the small green box in Fig3.5a. Note that of the three small upward deviations in the iGluSnFR signal after 29.2 s, only the second was counted as an event. **c)** A histogram of event amplitudes from active zone 1. The black trace is a fitted sum of six Gaussians.

3.2.7. Amplitude Clustering and Quantal Time Series

Based on the evidence that glutamate transients of varying amplitude were integer multiples of a unitary event or quantum (see **Fig3.6c**), I partitioned events into numbers of quanta using a Gaussian Mixture Model (GMM). Under this framework, the probability **p** of a value **x** (here referring to a sample observed peak deconvolution value) from a normal distribution with parameters μ and σ is given by:

$$p(x) = \sum_{i=1}^{K} \phi_i \mathcal{N}\left(x \middle| \mu_i, \sigma_i^2\right) \tag{7}$$

where **N** is the normal probability density function and **Φ** represents the mixing probability and sums to one, and **k** is the number of latent clusters. Note that depending on formulation, this can be equivalently defined as a combination of a Hidden Markov Model (HMM) (with cluster occupancy representing hidden states) and an Infinite Mixture Model (IMM) (as there are theoretically infinite different possible clusters). Several algorithms for clustering were tested and I found that Expectation-Maximization (EM) (Dempster, Laird et al. 1977) provided a quick and efficient approach. For each synapse, the algorithm was run up to 15 times, each with a different number of components and with the mean and variance parameters initialized randomly. The algorithm was iterated until acceptable convergence had been reached (defined by a user-set threshold). Following completion of the 15 runs, the output clusters were plotted with a histogram of the extracted events to allow the user to select the optimal partition based on these outputs, as well as a plot of data likelihood for each run. The user was then able to select the value for **k** and the corresponding partition of the data set, thus defining the vector **AQ** of estimated number of quanta for each event, an example of which is shown in **Fig3.2e**. Defining a time series as the number of quanta within each event allowed for computation of vesicle release rates and information theoretic measures.

3.3. Results

3.3.1. Testing of Time Resolution

The detection of events relies upon finding local maxima in the deconvolved traces. Noise within our records creates the potential problem of mislabeling two individual events as a single, higher quantal event (i.e., two uniquantal events occurring in close succession could be misclassified as a single 2-quantal event), as is shown by the simulation in **Fig3.7a**. In order to assess this 'temporal discrimination window' describing *the minimum time between events required before events can be reliably discriminated*, I first measured the signal-to-noise ratios (SNR) within our experimental data and then simulated a series of pulses of increasing inter-event intervals with matching SNR values. I then ran each of these simulations through the analysis sequence described above to estimate the temporal discrimination window.

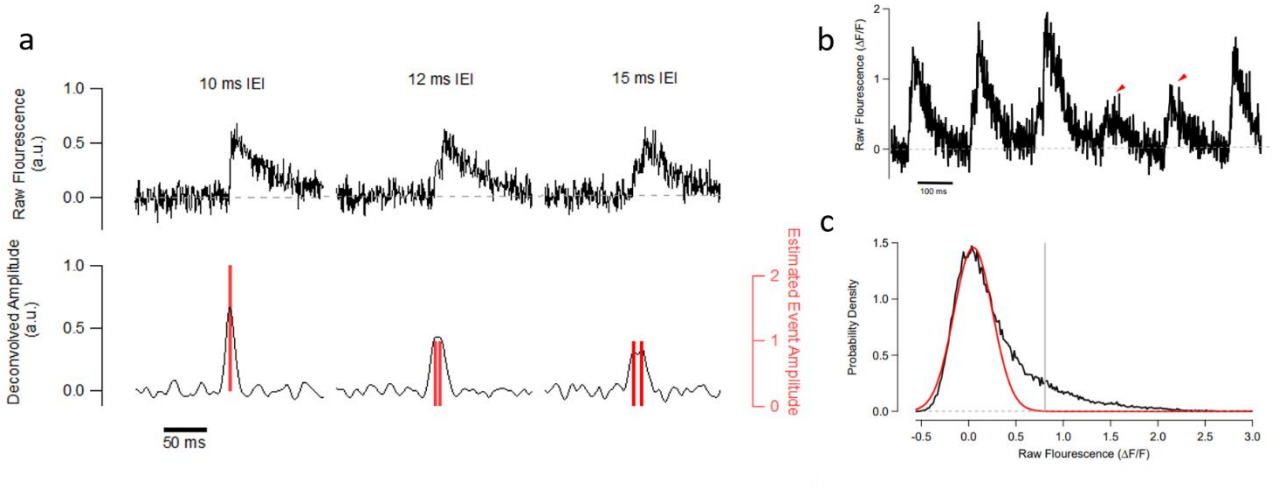

Figure 3.7: **Estimating the SNR within a recording.**

a) Left: Two uniquantal events separated by 10 ms. Note that at this IEI, these events are incorrectly classified as a single 2-quanta event (bottom). Middle: Two events separated by 12 ms generate two distinguishable maxima in the deconvolved trace counted as two distinct events, the amplitude and timing of which is shown by the vertical red bars (bottom). Right: Two uniquantal events separated by 15 ms are distinguished relatively easily. In all these simulations, the SNR was 4.05. **b)** Representative raw trace in response to a 5 Hz sinusoidal stimulus. The SNR value calculated for this terminal was 3.95. Note that red arrows show uniquantal events. **c)** Distribution of fluorescence values from the above recording. The red line indicates the Gaussian fit with mean 0.04, and amplitude 0.74. The black line shows the experimental data. The SNR is calculated by dividing the mean uniquantal amplitude (ΔF/F = 0.8, highlighted by gray line) by the standard deviation of the noise (0.2), represented by the Gaussian. Example of simulations used to estimate the temporal discrimination window.

The SNR within a recording was defined as the average amplitude of a uniquantal event divided by the standard deviation of the baseline noise. To compute the standard deviation of the noise signal I plotted the distribution of fluorescence values of the raw ΔF/F trace, found the first peak and then fit all values to the left of this peak (thus removing any possible contamination by signal) with a Gaussian, as shown in **Fig3.7b,c**. Note that this is not the same trace used to compute the threshold for event detection: that threshold was applied after deconvolution (Section 1.5). In a sample of 10 synapses, the SNR estimated in different traces ranged from three to eight, with the large majority greater than four. In 100 simulations using SNRs ranging from 3 to 8 (**Fig3.7a**), a SNR of four provided a temporal discrimination window of 10-15 ms (**Fig3.8**).

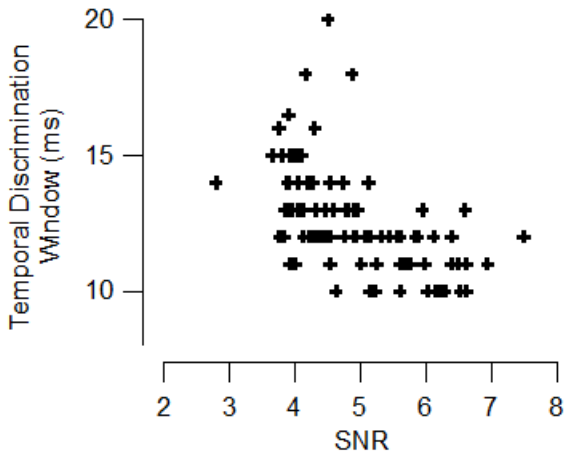

Figure 3.8: Scatter plot of temporal discrimination windows vs SNR values.

These simulations suggest that in low noise conditions (SNR > 5) we can reliably distinguish events separated in time by 10-15 ms. The mean temporal discrimination window value is 12.6 ms.

Given these clear disadvantages, how useful is this method? While MVR, defined on the electrophysiological timescales, will clearly be observed, how likely am I to collapse uni-quantal events into MVR events? Previous reports of quantal release have suggested that the synapse has a 'quantal' absolute refractory period, a period of time following the release of a vesicles where no further vesicles are released, measured to be at least 5 ms (Stevens and Wang 1995, Hjelmstad, Nicoll et al. 1997). Note that no precise mechanism has been identified for the observation, although it could be a result of either vesicle depletion or synaptic acidification by

vesicular protons. Thus, the biology of the synapse might be beneficial towards our approach – if vesicles do obey a refractory law, then it might not be as necessary to decrease our temporal discrimination window to levels attainable in electrophysiological recordings. As long as our window can approach the quantal refractory period, then we can be confident we are not incorrectly misclassifying uni-quantal events as multi-quantal events.

3.3.2. Testing for Linearity

One drawback imaging has relative to more conventional direct electrophysiological or electrochemical techniques is the issue of linearity. Due to the kinetics of the reporter used, the recorded fluorescence signal may not linearly indicate glutamate concentration; a DF/F value of 2 might not necessarily – and in fact rarely does - indicate twice the concentration of a molecule as a DF/F value of 1. Notably, saturation of the reporter can become an issue with optical reporters – if a substantial fraction of the available reporter used is bound to the molecule of interest at a given time, the signal will become saturated, limiting the maximal amount of fluorescence observed. As in this work I aim to develop a quantitative manner in which to count vesicles within glutamatergic events, the saturation of the iGluSnFR reporter is a particular concern.

The dissociation constant (Kd) of the iGluSnFR variant used here is ~4 µm (Marvin, Borghuis et al. 2013). As glutamate transients within the synapse are capable of reaching millimolar concentrations after vesicular fusion, it is possible that the iGluSnFR signal would saturate to a single vesicle, thus producing increasingly inaccurate reflections of quantal content as the signal increases. However, it is worth noting the slew of biophysical aspects involved in glutamate release. Saturation and linearity of a local iGluSnFR signal will depend upon on- and off-rates of glutamate-iGluSnFR interactions, diffusion properties of the reporter, spatial distributions of both iGluSnFR molecules, and vesicular transporters. Thus, an analytical solution to whether or not the iGluSnFR signal will saturate is a difficult question to answer.

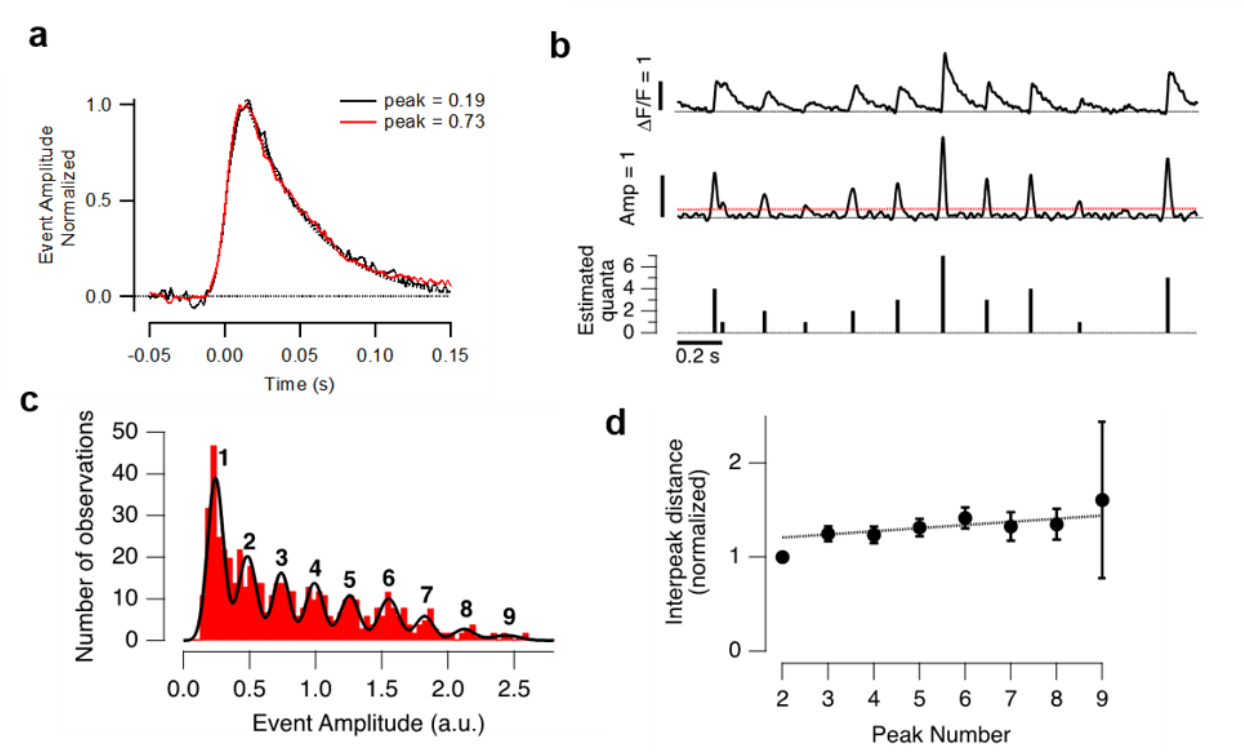

Figure 3.9: Experimental evidence of the linearity of iGluSnFR.

a) The average of 15 iGluSnFR events with a peak amplitude of DF/F = 0.198 ± 0.001 (black; mean ± sd) superimposed on the average of 15 events approximately four times as large (red; peak □F/F = 0.73 ± 0.01). After normalization, the small and large events superimpose. The decline from the peak occurs with a time-constant of 44 ms (dashed line). **b)** An example of the basic steps in the analysis by which events were counted and quantified. The raw iGluSnFR signal is shown at top (stimulus at 5 Hz, 80% contrast; Savitsky-Golay filter 21 ms). The middle trace shows the results of Wiener deconvolution using a kernel of unitary area and the shape shown in **a.** The time and amplitude of each event was obtained from the local maxima above a threshold (dashed red line). Similar analysis could be carried out in 150 independent experiments. **c)** A histogram of event amplitudes for the active zone featured in **b** (n = 547 events accumulated using stimulus contrasts of 30%, 80% and 100%). The black line is a fit of nine Gaussians, identified using a Gaussian Mixture Model. Note that the variance of successive Gaussians did not increase in proportion to the peak number. The first peak had a value of 0.24 and the distance between peaks averaged 0.26, indicating the existence of a quantal event equivalent to ~0.25. The amplitude of the quantal event averaged 0.23 ± 0.01 (mean ± sem, n = 20 synapses from independent experiments). **d)** The mean distance between successive peaks (normalized to the distance between the first and second peaks) plotted as a function of the peak number. Collected results from n = 6 synapses. Points show mean ± sem. The dashed line is a linear fit with a slope of 0.03 ± 0.12 (mean ± sem), which is not significantly different from zero. There were no signs of saturation of the iGluSnFR signal for events composed of up to 9 quanta.

Given the uncertainties in modelling iGluSnFR signals, I took an experimental approach to assessing whether iGluSnFR transients might begin to saturate in response to larger MVR events. To begin, I identified and collected isolated fluorescence events from what the algorithm identified as unitary or 4-quantal events (**Fig3.9b)** and averaged them. As seen in **Fig3.9a**, these events had identical shape, indicating that at this level no saturation had occurred. Amplitude histograms from individual active zones were constructed using stimuli of both low contrast, when the distribution of event amplitudes is shifted towards smaller numbers of quanta, and high contrast, when there are more large events. An example of such a histogram is shown in **Fig3.9c**, where eight distinct peaks are evident. I then measured the interpeak distances from a sum of Gaussians fit and asked whether the distance between peaks might be reduced for events containing larger numbers of quanta, as would be expected if there was significant saturation of the reporter. Collected results from six active zones are shown in **Fig3.9d**. The change in interpeak distance was not significantly different from zero up to events composed of 9 quanta, indicating almost perfect linearity over this range. In comparison, the largest events I observed from a sample of 51 synapses were equivalent to 11 quanta. It therefore seems unlikely that my estimates of quantal number were skewed by saturation of the iGluSnFR reporter. More recently, evidence from mouse cortex also indicates iGluSnFR's linearity, confirming our results (ref).

3.3.3. Isolating Single Ribbons

The point-spread function (psf) of the microscope had a $FWHM_{xy}$ of 0.7 μm in the x-y, allowing active zones separated by 1 μm to be easily distinguished in the iGluSnFR signal along a single linescan (**Fig3.2**). The psf in the *z* dimension was, however, significantly larger ($FWHM_z$ = 2.2 μm), raising the possibility that two or more active zones at different *z* depths might be considered as one if they coincided closely enough in the *x* dimension of the linescan.

To remove the possibility of recording signals from adjacent terminals, I used zebrafish in which only a fraction of bipolar cells were expressing iGluSnFR and chose terminals which were spatially isolated from others expressing the reporter, as shown in **Fig1a** and **Fig3.10**. To assess the probability of conflating signals from different active zones *within* one terminal I measured the numbers and distribution of synaptic ribbons that holds vesicles close to the sites of fusion and which colocalize with calcium channels (Zenisek, Horst et al. 2004, Lagnado and

Schmitz 2015). Ribbons within individual terminals were visualized in fixed samples using an antibody to ribeye, as shown in **Fig3.10a** and **b**. The zebrafish larvae used for these measurements were prepared in the same way as those used for measuring glutamate release. The ImageJ tool DiAna, was used for object-based 3D co-localization and distance analysis(Gilles, Dos Santos et al. 2017). The distance between ribbons was measured between their centers of mass, but I did not count "floating" ribbons, defined as those that were more than 0.5 mm from the surface membrane (Euclidean distance). Floating ribbons that are not attached to the surface are a common feature of bipolar cells (Zenisek, Horst et al. 2004) and constituted 28% of 58 ribbons in a sample of 4 terminals. The density of ribbons attached to the surface membrane averaged 0.16 mm^{-2}, which was similar to a previous measurement of 0.12 ribbons mm^{-2} made in regions of flattened membrane in bipolar cells from goldfish (Zenisek, Horst et al. 2004). The distance between nearest ribbons averaged 0.96 ± 0.4 μm, and the distribution of values is shown in **Fig3.11**.

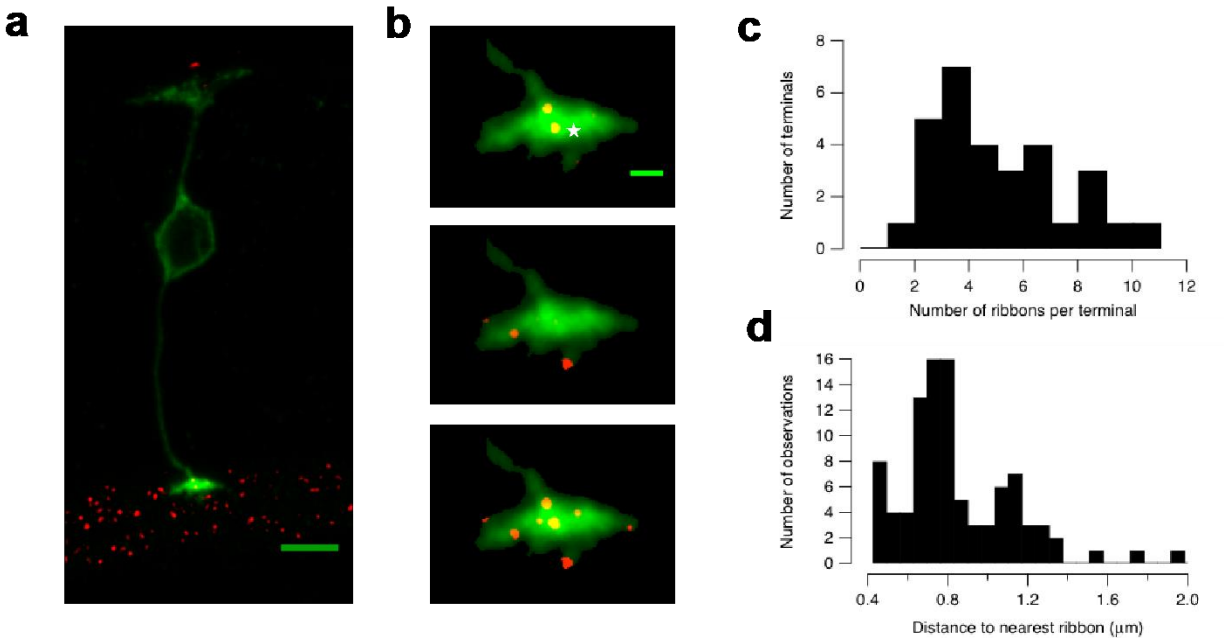

Figure 3.10: Counting ribbons.

a) Confocal image of a bipolar cell expressing iGluSnFR (green) and synaptic ribbons labelled with a ribeye B antibody (red). Note the ribbons scattered throughout the inner plexiform layer. Scale bar 5 □m. **b)** In all three panels, the green image is a maximum intensity projection of the iGluSnFR signal through a volume of 34.16 □m^3 containing the terminal of the cell shown in **a**. A threshold was then applied to mask out fluorescence beyond the terminal. Superimposed in the top panel is a maximum intensity projection of the red signal through three planes (each separated by 0.25 μm) centred towards one side of the

terminal. Two ribbons can be seen. The middle panel shows a similar projection of the red signal at a z distance of 3 μm from the first, where three ribbons can be seen. In the bottom panel, the red channel shows a maximum intensity projection through the entire volume of the terminal. Scale bar 1 □m (top panel).

To compute the probability of 'collapsing' the signal from two separate ribbons, I first calculated the lateral and axial resolutions of our microscope (ω_{xy} and ω_z) (Zipfel, Williams et al. 2003):

$$\omega_z = \frac{FWHM_z}{2\sqrt{ln2}} = 1.3\ \mu m \tag{8}$$

$$\omega_{xy} = \frac{FWHM_{xy}}{2\sqrt{ln2}} = 0.42\ \mu m \tag{9}$$

I then constructed a 'resolution' volume by creating an ellipsoid with major axis equal to the axial resolution and minor axes equal to the lateral resolution, given by equation:

$$\frac{x^2}{{\omega_{xy}}^2} + \frac{y^2}{{\omega_{xy}}^2} + \frac{z^2}{{\omega_z}^2} = 1 \tag{10}$$

Thus, a ribbon at the origin of this ellipsoid would not be discriminated from any other ribbon lying within this volume. A section through this "resolution volume" is illustrated in **Fig3.11a**, on which is superimposed "nearest neighbor volume" defined by a second ribbon at a distance "m" from the first, in any direction. Assuming that all ribbons are randomly distributed relative to each other, the distribution of ribbons as a function of "m" is uniform over the surface of the sphere defined by radius "m". The probability of collapsing two ribbons at distance "m" can then be calculated as the surface area of the intersection of the two volumes divided by the surface area of the sphere. As the resolution ellipsoid defined in equation 10 is in fact a prolate spheroid, this value can be computed analytically or numerically using spherical caps, and the **Fig3.11b** shows the probabilities of collapsing two ribbons as a function of the distance "m" between nearest ribbons. For the sample of 4 terminals and 42 non-floating ribbons, I compute

an average probability of collapsing two nearest ribbons as 8%. Statistical analyses show equivalent results: as the probability of a signal being collapsed as a function of the distance between cells is here (rather ironically) a deterministic function, viewing the distance between ribbons as a Gamma-distributed random variable LOTUS can be applied, yielding: $E[Collapsed(R)] = \int_0^\infty p(r)Collapsed(r)dr$.

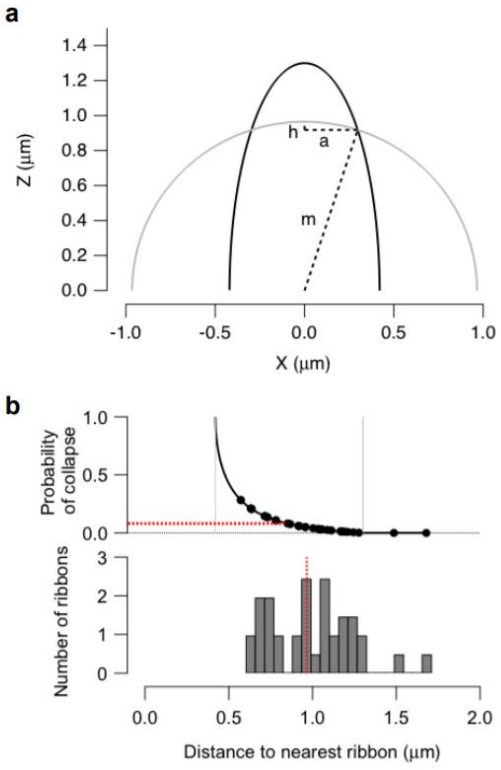

Figure 3.11: Model for calculating the probability of conflating signals from two ribbons.

a) Example sphere (gray line) and ellipsoid volumes (black line). The spherical cap of their intersection is defined from the radius of the sphere **m** (using the average nearest distance between ribbons), the circle radius at the z-value of the two volumes intersection **a**, and the height of the cap **h**. **b)** Top) The probability of collapse as a function of m (black line) overlayed with the values computed from the data (black markers). The mean probability is 8% (gray dashed line). b, Bottom), the distribution of distances between ribbons. Gray dashed line indicates mean nearest distance of 0.96 mm.

3.4. Discussion

In recent the past two decades, a great deal of work has focused on the phenomenon of MVR (Korn, Sur et al. 1994, Tong and Jahr 1994, Singer, Lassova et al. 2004, Foster, Crowley

et al. 2005, Huang, Bao et al. 2010, Chamberland, Evstratova et al. 2014, Rudolph, Tsai et al. 2015, Vaden, Banumurthy et al. 2019). However, in many of these studies, the underlying tissue being studied is damaged during the data acquisition process – as can be seen in works utilizing electrophysiological patch clamp or amperometry. It has thus been difficult to attribute any functional significance to the process. Here, I developed an analysis framework that overcomes many of these challenges. Using 2p microscopy in conjunction with the iGluSnFR reporter, the experiments can be undertaken in the intact zebrafish larvae, guaranteeing the connections of the circuit remain intact, while still retaining the signal-to-noise required for decomposition of the signal into units of individual quanta. While iGluSnFR has been used to observe glutamate release in various organisms and cell types, this is the first work in my knowledge describes a method for the quantal decompositions found in electrophysiological works using 2p microscopy in the intact organism.

3.4.1. Relation to Previous Work

Analyses attempting to quantify vesicular release into units of induvial quanta are not novel. In fact, this decomposition has a long history in both vision and hearing. However, the two main techniques used in these analyses – either the patch clamp analysis of miniature excitatory post-synaptic potentials (EPSCs), or the analysis of amperometric changes caused by a cell's loss of synaptic membrane during vesicle exocytosis – in general require more invasive measures - which almost always cause damage to the underlying circuitry being analyzed. In contrast, the more recent work on using iGluSnFR as a reporter in the nervous system has enabled researchers to interrogate aspects of synaptic transmission in completely intact circuits. With a far greater SNR than calcium reporters, iGluSnFR has been successfully used to monitor the glutamate released across a large population of cells in both the mouse retina and cortex (Franke, Berens et al. 2017, Baden, Euler et al. 2020, Oesterle, Behrens et al. 2020), as well as in nearly all regions of the zebrafish brain (Vladimirov, Wang et al. 2018, Ahrens 2019). However, these experiments often focus on a coarser resolution than used in the present work, as a notable benefit of 2p imaging is its ability to record population data. Here, rather than imaging across a wider spatial window, I focused tightly on a single axon terminal, attempting to maximize the temporal resolution of the signal at a great cost to spatial resolution. While no individual step of the analysis is new in neuroscience, the combination of each step with the data made available by the iGluSnFR signal allows for an optical approach to the techniques that were previously only achievable with electrophysiology.

3.4.2. Limitations – computational and physical

There are two main limitations to the described method – the physical limitations of the reporter and imaging apparatus, as well as the computational limits. Here, we are utilizing 128 pixel linescans in order to achieve the temporal resolution required to decompose glutamatergic events. While the imaging speed can be increased by the use of a resonance scanner, note that the kinetics of the reporter will remain unaffected, with a 1 ms rise time. Increasing the acquisition rate of the imaging system then may slightly increase our ability to decompose events, but most likely not significantly – Wiener deconvolution still relies upon differences in the spectral content of the reporter signal vs noise, and increasing the imaging frequency will not affect noise distribution. Another constraint to our analysis, rather than being aided by the use of linescans, detracts from it – photobleaching. As we are here using linescans to record glutamatergic activity, all laser intensity is focused on a small strip of tissue (5-10 microns). While focusing all laser power on an isolated area like this can increase signal-to-noise ratios, it also can damage the reporter, decreasing the proportion of available reporter in time and reducing in a gradual decrease of signal intensity. While some of this can be corrected for, bleaching does reduce the length of time which a single BC can be recorded, limiting the experimental data collected from any cell – and thus disallowing certain analyses that require large amounts of data.

How sure of our decomposition can we be? In our analysis, we cluster each detected glutamatergic event into units corresponding to the number of vesicles released in that event. As a first pass, we can think about how many different possible ways we could cluster our data. In combinatorics this is referred to as all possible partitions of a set, and the number of distinct ways in which one can partition a set of N objects is given by Bell Number. Notably, Bell Number rises tremendously quickly: a set of 13 object can be partitioned in over 4 million different ways. As such, for a reasonably-lengthed recording, it is not possible to directly determine the probability of each possible partition in order to find the maximum – even with a supercomputer it would take centuries to accomplish this task for a large dataset. Due to this, distinct algorithms have been created to tackle this problem. Here, we used the Expectation-Maximization (EM) algorithm for its efficiency, although this is perhaps a less effective algorithm than others. For this reason, I also tested a different algorithm – a Gibbs Sampler under the framework of a combination of a Hidden Markov Model (HMM) and an Infinite Mixture Model (IMM). This algorithm, while taking considerably longer to run than the EM algorithm, is guaranteed to converge to the true underlying distribution with enough iterations. A comparison of the results of either algorithm revealed similar results with respect to the models Maximum a Prior estimate (MAP), and the EM algorithm was consequently chosen for efficiency.

3.4.3: Divergence from Convolutional Statistics

If we consider the single vesicle as the most basic quantal element of vesicle release, given by a distribution with a finite mean μ_1 and variance σ_1^2, then the distribution of the sum of individual vesicles should be defined by convolutional statistics (which represent the sum of random variables). Notably, equation 7, which has a total of 2k free parameters, could be reduced to

$$p(x) = \sum_{i=1}^{K} \phi_i \mathcal{N}(x \big| k\mu_1, k\sigma_1^2),$$

with a total of 2 free parameters. Here the distribution corresponding to k simultaneously released vesicles should have mean $k\mu_1$ and variance $k\sigma_1^2$. While our estimates did indeed show that the mean of each distribution was a multiple of the single vesicle distribution, we did not see this trend in variances – the variances for each sequential cluster did not increase as much as we would have expected. While it is uncertain exactly why this is the case, I do note that the signal analyzed was pre-processed before being fed into the mixture model. It is possible that the Wiener deconvolution in particular affected this result. While it does attempt optimal denoising of the signal, overlapping frequency spectra of noise vs. impulse kernel can disproportionately affect lower amplitude signals with lower SNR, inflating their variance relative to higher quantal and SNR events.

Chapter 4: Experimental Analysis of MVR

4.1. Introduction

Throughout a large portion of the history of neuroscience, the study of information transmission has largely been synonymous with the study of action potentials. In the prototypical neuron, synaptic information is integrated in time and space (from multiple inputs), resulting in membrane depolarization. Once the membrane voltage potential reaches a threshold, the neuron 'fires' and releases all of its neurotransmitter onto postsynaptic cells, undergoes a short refractory period where the cell remains inactive, after which the process repeats. Due to the very nature of the action potential, this firing is binary: at any given time, the output of a neuron consists of either a spike or no spike, corresponding to a binary system with the symbols 0 and 1. Within this framework of small time windows, information transmission can then be naturally studied via binary, digital processes.

A great deal of understanding of neural information transmission has been gained by using this framework. Many aspects of neural activity can be described in this manner with reasonable efficiency and understanding. Using a binary system allows not only for simple computation of information theoretic measures (Strong, de Ruyter van Steveninck et al. 1998), it also allows for more complex representations of neural responsivity, for example the spike triggered average (STA), to be easily calculated (Schwartz, Pillow et al. 2006). The times at which this binary signal spikes can even be modelled simply using Poisson Processes (PPs) (Johnson 1996). It even lends itself to the pop-scientific metaphor of the brain as a computer.

However, this binary framework does fail, for multiple reasons. For one, not all cells are capable of action potentials – many cells, especially in the early sensory systems, operate by adjusting their membrane potential in a graded or analogue fashion. Thus, information in these cells is not represented by binary activity, but rather a continuous value (de Ruyter van Steveninck and S.B. 1996). The second failure of the binary view of information transmission is the phenomenon of multivesicular release (MVR), wherein multiple vesicles are released nearly simultaneously (Auger, Kondo et al. 1998, Glowatzki and Fuchs 2002, Singer, Lassova et al. 2004, Christie and Jahr 2006, Higley, Soler-Llavina et al. 2009, Huang, Bao et al. 2010). Notably, cells equipped with what is referred to as the synaptic ribbon exhibit both of these properties – they grade their membrane potential in a largely continuous, analogue fashion, and they release multiple vesicles nearly simultaneously, utilizing MVR (Parsons and Sterling 2003). Electrophysiological evidence from these synapses suggests that these MVR events can be

coordinated with up to microsecond timescales (Singer, Lassova et al. 2004), far faster than the timescales of voltage leak within postsynaptic neurons.

While previous work has indeed extended the traditional binary analysis of action potentials to multiple signals by increasing the length of the time window in which signals are analyzed, little of this theoretical analysis has spilled over into the field of synaptic transmission. Using windows of up to 100 ms, it has been shown that spike bursts – quick successions of action potentials tightly grouped in time – can transmit information in using multiple symbols (here corresponding to the number of spikes in a burst) (Zeldenrust, Chameau et al. 2013, Carrillo-Medina and Latorre 2018, Zeldenrust, Chameau et al. 2018, Zeldenrust, Wadman et al. 2018). However, analyses of this sort do not provide evidence for the nervous systems use of multiple symbols in synaptic transmission. For example, in many brain regions the number of vesicles released per spike can be far less than one (Lisman, Raghavachari et al. 2007). In these cases, it is unlikely that multiple spikes occurring in a burst would reliably release multiple vesicles that would allow for effective summation of the synaptic signal into multiple symbols useful for information transmission. Here, by directly observing the release of multiple quanta of neurotransmitter in windows of as low as 10 ms, we explore how the nervous system might use multiple synaptic symbols for information transmission.

The transfer of information between many cells, such as cells equipped with ribbon synapses, is not a binary signal, as was once believed, but rather a ‘multi-nary’ signal, where the symbols consist not only of zero and one, but also include two, three, &c. This arises from the fact that quanta released in MVR events can be coordinated to microsecond timescales. A cell that releases events consisting of one to eleven vesicles produces twelve symbols – one for each quantal event type (the number of vesicles/quanta in an event, here from one to eleven) and one for zero vesicles. Here, it is necessary to note that although the system allows for multiple symbols to be used, it does not necessarily indicate that they *are* being used to transmit information. For example, it may simply be the case that the system conveys information by either releasing all available vesicles or none. Then, the number of vesicles in an event conveys no information – only their presence. Here, note that may be how central synapses operate – all vesicles available vesicles are released during a spike, and the number of vesicles released in an event corresponds to the adaptive state of the neuron. What we would ideally like to show, then, is that the number of vesicles in an event is utilized in an informative way, i.e., are they modulated by some feature of the visual environment.

While MVR has been shown to occur in diverse brain regions (Auger, Kondo et al. 1998, Wadiche and Jahr 2001, Christie and Jahr 2006, Higley, Soler-Llavina et al. 2009, Huang, Bao et al. 2010), here we focus our attention on the retina. I show MVR is utilized in an informative manner to transmit information in the retina by constructing an 'amplitude' code, wherein information pertaining to visual contrast is encoded by modulating the number of vesicles in a glutamatergic event. By imaging glutamate release *in vivo* from axon terminals of Bipolar Cells (BCs) in the zebrafish larvae using two-photon microscopy, I can estimate the number of vesicles in a glutamatergic event (see Chapter 3). Stimulating the retina with full-field contrast modulated light, we observe that the distribution of vesicles per event is contrast-dependent, indicating that MVR is able to transmit contrast information. Information theoretic analyses of the resulting responses indicate additionally that not only are MVR events capable of conveying information, but they convey more information per vesicle than uniquantal events, making it possible that MVR increases the efficiency of information transmission in the retina.

4.2. Methods

4.2.1. Transmitter Triggered Average (TTA)

The Transmitter Triggered Average (TTA), first proposed by Abbot and Regehr (Abbott and Regehr 2004), can be interpreted as an extension of the Spike Triggered Average (STA) to more than one types of neural symbol. In the traditional spiking neuron, there is only one distinct active symbol – the spike. The STA then recovers a single filter. Here, we find not a single active symbol, but multiple (corresponding to uni-quantal events, two-quantal events, etc.). The TTA recovers a filter for each event type; while the STA is a single vector for the spikes filter, the TTA consists of a matrix – one vector for each observed event type. The TTA can be described by the equation:

$$\overrightarrow{s_q} = \frac{1}{n_q}\sum_{i=1}^{n_q}\overrightarrow{s_i}(t)$$

Where n_q is the number of q-quantal events, $\overrightarrow{s_i}(t)$ is the stimulus preceding q-quantal event i, and $\overrightarrow{s_q}$ is the recovered filter for all q-quantal events. While the analysis is novel in the context of vesicular events, it is noteworthy that a comparable analysis exists for the case of spike bursts (Zeldenrust, Chameau et al. 2013, Carrillo-Medina and Latorre 2018) wherein multiple action potentials occur in a quick, burst-like succession. As the TTA extends the range of symbols to

beyond zero or one, so does the analogous event-triggered average for either singular spikes, or groups of spikes occurring in short succession (de Ruyter van Steveninck, Bialek et al. 1988, Zeldenrust, Chameau et al. 2018, Zeldenrust, Wadman et al. 2018).

4.2.2. Information Theory

To measure the specific information (DeWeese and Meister 1999) between the distribution of stimuli **S** and the observance of a single quantal event type **q**, we first divided the responses into 20 ms bins, such that no two events can be found in any bin. Note that here, as in classical notation, we refer to **Q** as the random variable, and **q** as a single possible occurrence from that distribution. The probability of observing a given quantal event type is then calculated empirically as the number of bins containing an event of that type divided by the total number of bins for each stimulus. Note that here only eleven stimuli were used, and their distribution was experimentally set to be uniform.

This results in an empirical estimate of joint probability mass function p(Q,S). We find the conditional distribution p(S|Q) by the definition of conditional probability: $p(S|Q) = \frac{p(S,Q)}{p(Q)}$. From these distributions we can the compute the specific information as the difference between the stimulus entropy and the conditional entropy of the stimulus given a particular symbol:

$$\boldsymbol{I_2(S,q) = H(S) - H(S|q)}$$

$$\boldsymbol{= -\sum_{s \in S} p(s)\, log\, p(s) + \sum_{s \in S} p(s|q)\, log\, p(s|q)}$$

4.3. Results

4.3.1. MVR is contrast dependent

How might MVR be utilized to transmit information across the synapse? One of the most basic properties of the visual scene a BC responds to is temporal contrast; I thus stimulated individual axon terminals with a light stimulus consisting of a sinusoid of varying temporal contrast while recording their glutamatergic release. To do so, I located isolated axon terminals expressing the iGluSnFR reporter as in **Fig4.1a** and drew lines consisting of 128 pixels across their terminals, as in **Fig4.1b**. Here, the mosaic expression of the BCs proved useful – not all cells express the reporter, and this allows me to identify isolated cells to avoid potentially collapsing signal from multiple cells in the z-direction. While reducing the recording to a single

line of 128 pixels (**Fig4.1b)**, this allowed me to scan each point with a temporal frequency of 1 kHz – enough to enable decomposition of glutamatergic events into units of vesicular quanta – the events amplitude.

To begin, I stimulated axon terminals of isolated BCs with a full-field 5 Hz sinusoidal at 100% contrast and recorded the resulting fluorescence, as in **Fig4.1c**. Note the heterogeneity of the amplitudes – rather than operating with a fixed amplitude as most action potentials do, the glutamatergic output seems to vary. Moreover, the observed amplitudes correspond to units of individual vesicles (see chapter 3). Thus, this contrast evokes MVR events of variable amplitude. Stimulating the BC thus evokes the release of glutamatergic events consisting of variable numbers of vesicular neurotransmitter – different quantal events. In order to understand how this might reflect information transmission, I stimulated the cell at varying contrasts.

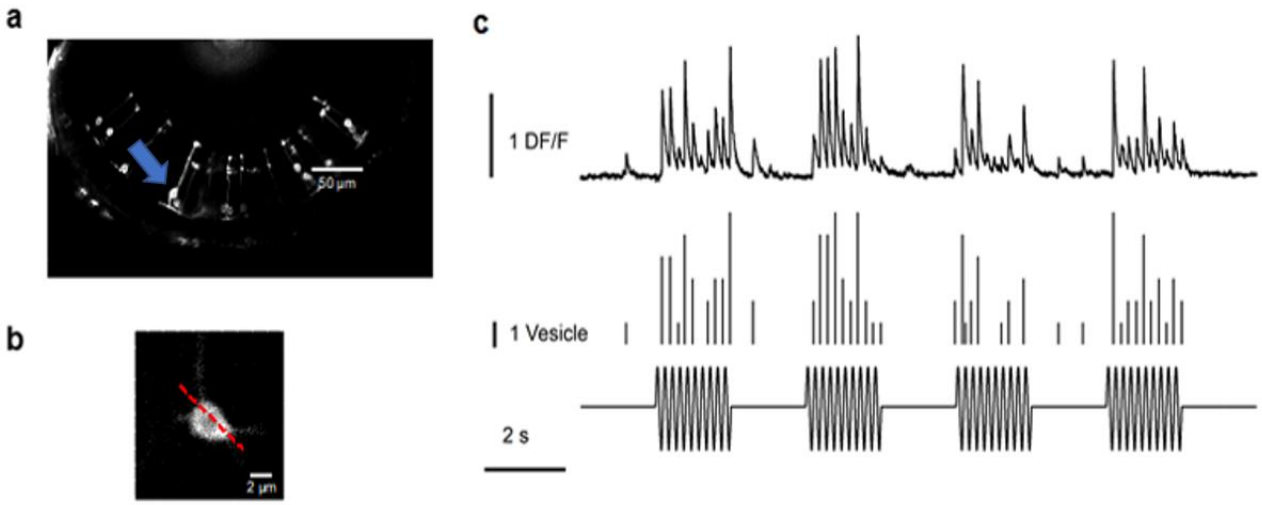

Figure 4.1: Measuring glutamate release from zebrafish BC axon terminals *in vivo*.

a) The zebrafish retina under a 2p microscope, showing the mosaic expression of the iGluSnFR reporter over several BCs. Blue arrow shows an individual BC, pointing towards the cell body. b) An individual BC axon terminal, with the recording site for a linescan shown. c) Top: recorded response to a 5 Hz 100% contrast sinusoidal wave (bottom). Middle: The estimated number of quanta composing each recorded glutamatergic event

b)

Higher temporal contrasts regularly evoked higher quantal events, as shown in **Fig4.2a-c**, where I stimulated an AZ with a low (30%) contrast stimulus as well as a high (100%)

contrast stimulus. Note that not only does the rate of events increase with higher contrast, but also the distribution of vesicles in an event, as can be seen in **Fig4.2d** where I plotted the probability mass functions for events at low and high contrast. Note that at the higher contrast the distribution of quanta per event is shifted towards higher values, indicating that higher quantal events are preferentially released at higher contrasts. Here, we can begin to see how an amplitude code might be used to convey information – by shifting the distribution of event amplitudes as a function of contrast.

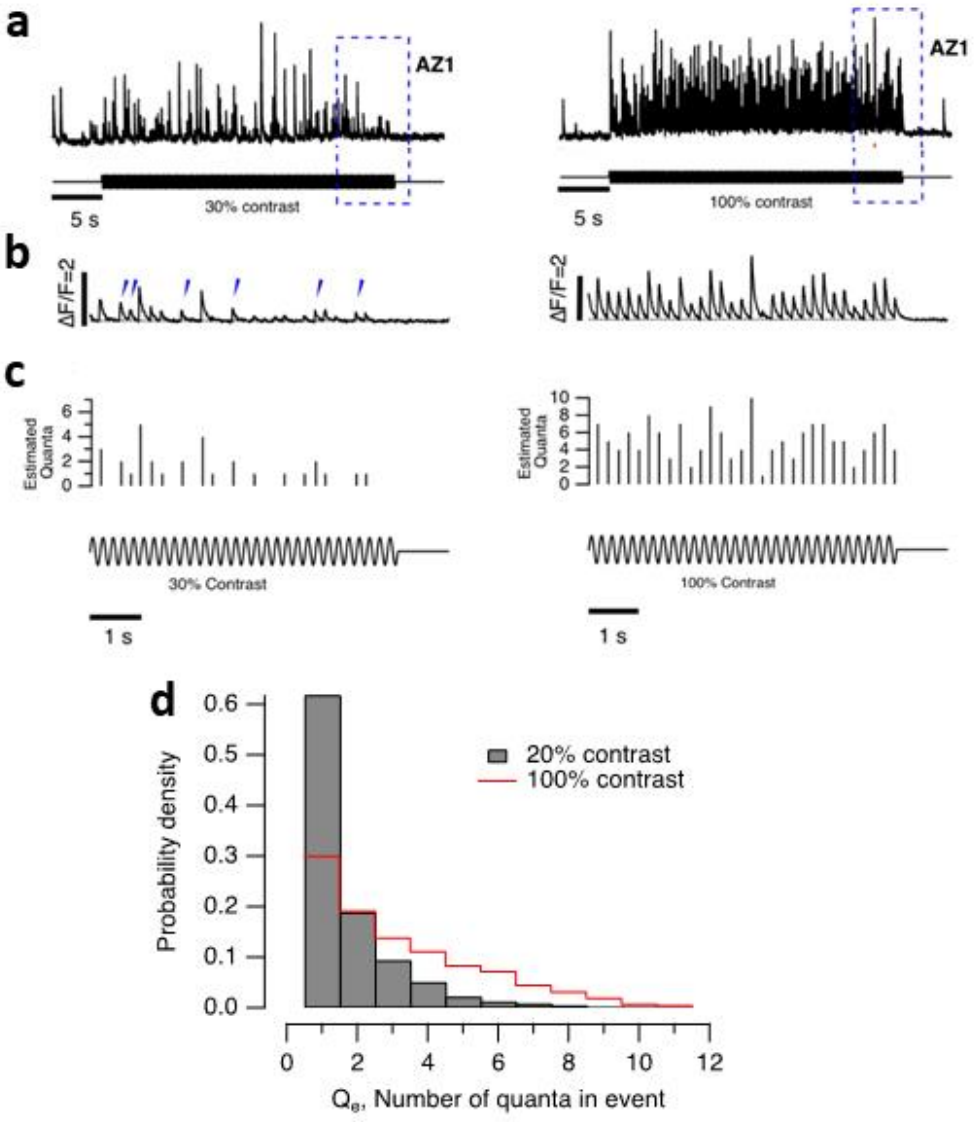

Figure 4.2: Contrast-Dependence of MVR.

a) Trace extracted for a single AZ showing changes in iGluSnFR fluorescence in a synaptic terminal in response to a full-field stimulus modulated at 5 Hz (sine wave). The left-hand side shows the responses of an active zone of a terminal stimulated at 30% contrast and the right-hand side the profile of the same terminal stimulated at 100% contrast. **b)** Expansion of the period shown boxed in b**. c)** Estimation of the number of quanta per event. The stimulation protocol is represented below. **d)** Probability mass functions for the quanta per event for a synapse presented with 5 Hz sine wave at 100% or 20% contrast. Note the shift in quanta per event at higher contrast.

4.3.2. Amplitude and Rate Coding Strategies

To further explore this contrast dependence, I first utilized a short stimulus consisting of a 5 Hz sin wave modulated from 0 to 100% in steps of 20%, as seen in **Fig4.3a.** A quick analysis of these data allows us to create a contrast response function where the response consists of the total number of released vesicles in a cycle of the stimulus, and then identify the half maximum point (taken to the nearest 10%) where the contrast response function is generally steepest, as shown in **Fig4.3b**. I then further stimulated these cells with contrasts plus or minus 10% from this half point in steps of 2%, as in **Fig4.3c.** This procedure allows us to explore the steepest area of a cell's response curve, regardless of where that region is located. From these data, I computed two other metrics – the average number of vesicles per event, and the average number of events per cycle. Here, we see that some terminals modulate the rate of vesicles releases more than the number of vesicles per event – these we say encode information predominately by rate coding (**Fig4.3d**). The other group modulates the number of vesicles in an event more strongly than it does the rate of events – and we consider these cells amplitude coded cells (**Fig4.3e**). Here, I refer to this first group as predominately 'rate-coders', while the latter group I refer to as predominately 'amplitude-coders'. Notably, many of the amplitude coding cells reached a saturation point in rate coding – once a certain event rate was reached, these cells encoded any further contrast solely by modulating event amplitude, and not rate.

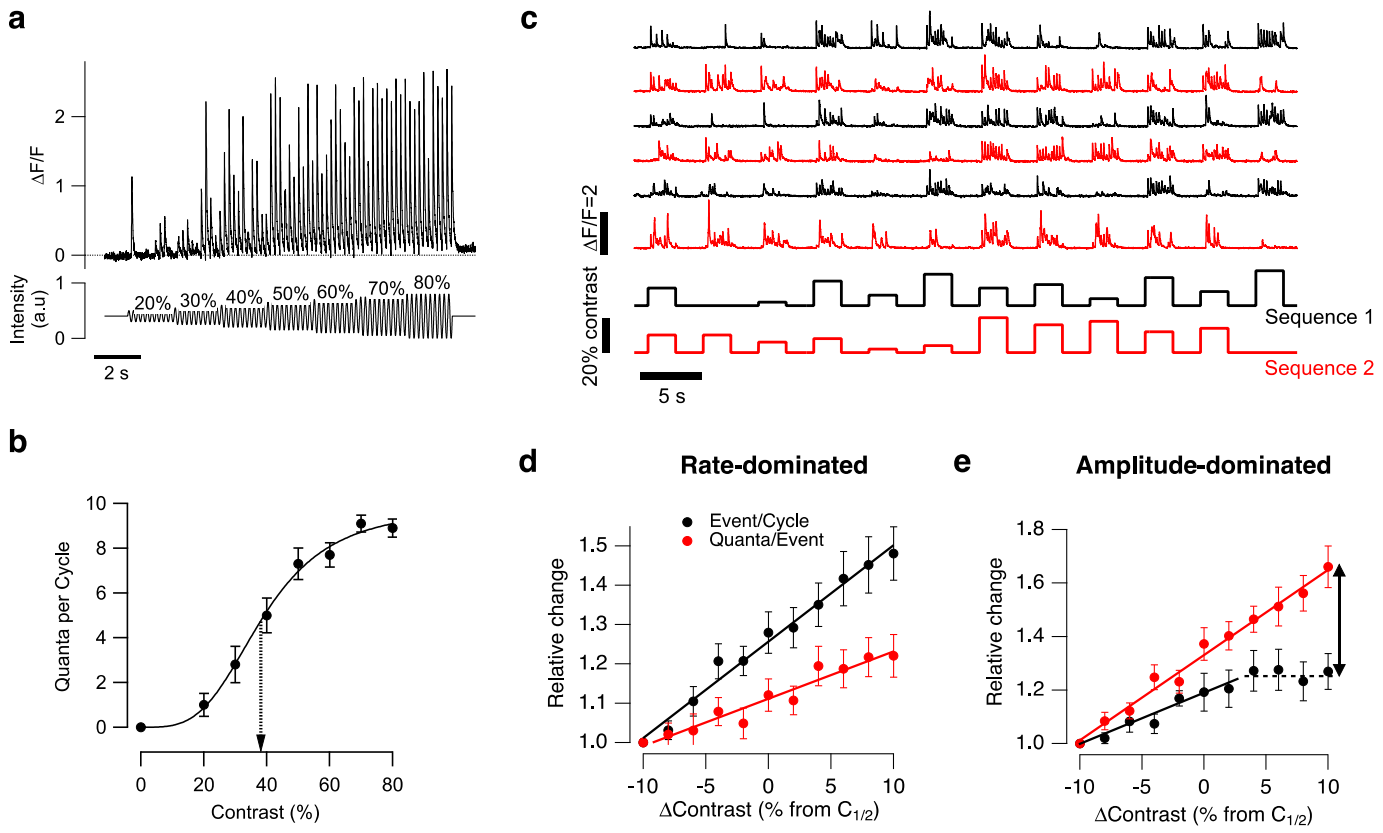

Figure 4.3: Demonstration of an amplitude code.

a) Example 5 Hz, 20-80% contrast stimulus used to isolate the most sensitive region of the contrast function. Note that both the frequency of events as well as the amplitude increases as a function of contrast. **b)** contrast response function from cell in a), arrow indicating half maximum, taken from 38 independent experiments. equation of the form $C = (R_{max})(C^h / C^h + C^h_{1/2})$, where $R_{max} = 9.8$ quanta per cycle (49 quanta s^{-1}), $h = 4.5$ and $C_{1/2} = 39\%$ (broken arrow). Each point shows the mean ± s.e.m. for n = 10 cycles of the stimulus. **c)** Stimulus set used for quantifying rate and amplitude information in 55 independent experiments. The contrasts spanned ±10% of the range around which the contrast sensitivity was highest in steps of 2%, all lasting 2 s and of a constant mean luminance and 5 Hz frequency. **d**) The relative change in synaptic activity around $C_{1/2}$. The average number of events per cycle (E_c) is compared with the average number of quanta per event (Q_e). Stimuli were delivered in 2-s episodes with a 2-s rest, as shown in Fig4.4a. In these n = 55 synapses, an increase in contrast caused the E_c to rise more steeply than the Q_e. **e**, In the remaining n = 17 synapses, the Q_e rose more steeply than the E_c, which then saturated (broken line) such that further increases in contrast were signaled only by increases in the number of quanta per event (vertical arrow). Points in **d** and **e** show the mean ± s.e.m.

Here, we note without any mathematical formalism that we can already see how MVR is transmitting information. If higher quantal events are more likely to be observed at higher contrasts, then observing a higher quantal event is a good indication that the stimulus that evoked this response was a higher contrast.

4.3.2. TTAs are contrast-dependent

If higher quantal events are more likely to be observed in response to higher contrasts, we would expect the average stimuli preceding a higher quantal event would be of higher contrast than a lower quantal event. To examine this, we utilized the Transmitter Triggered Average (TTA). Like the standard Spike Triggered Average, the TTA utilizes reverse correlation on a noise stimulus to estimate the average stimulus that precedes a spike – a simple way of quantitatively describing a cells linear receptive field (Dayan and Abbott 2001). Unlike the STA, though, the TTA has n different recovered filters – one for each quantal event type observed in the recording. In a STA, there is only one active symbol – a spike – and thus the STA consists of a single filter. In a TTA, there are multiple active symbols – one for the uni-quantal events, one for the two-quantal events,… - and thus the TTA consists of n filters (see **Fig4.4a** for an example TTA analysis).

If MVR truly is preferentially triggered by higher contrasts, then we would expect to see more contrast in the higher quantal filters than we do in the lower quantal filters, as indeed is the case. **Fig4.4b** shows an example of the TTAs recovered for a single cell, with increasing line darkness indicating higher quantal events. Note that the TTAs for higher quantal events (darker lines) for this active zone show a stronger decrease in light intensity preceding a spike (and thus higher contrast) than lower quantal events (lighter lines).

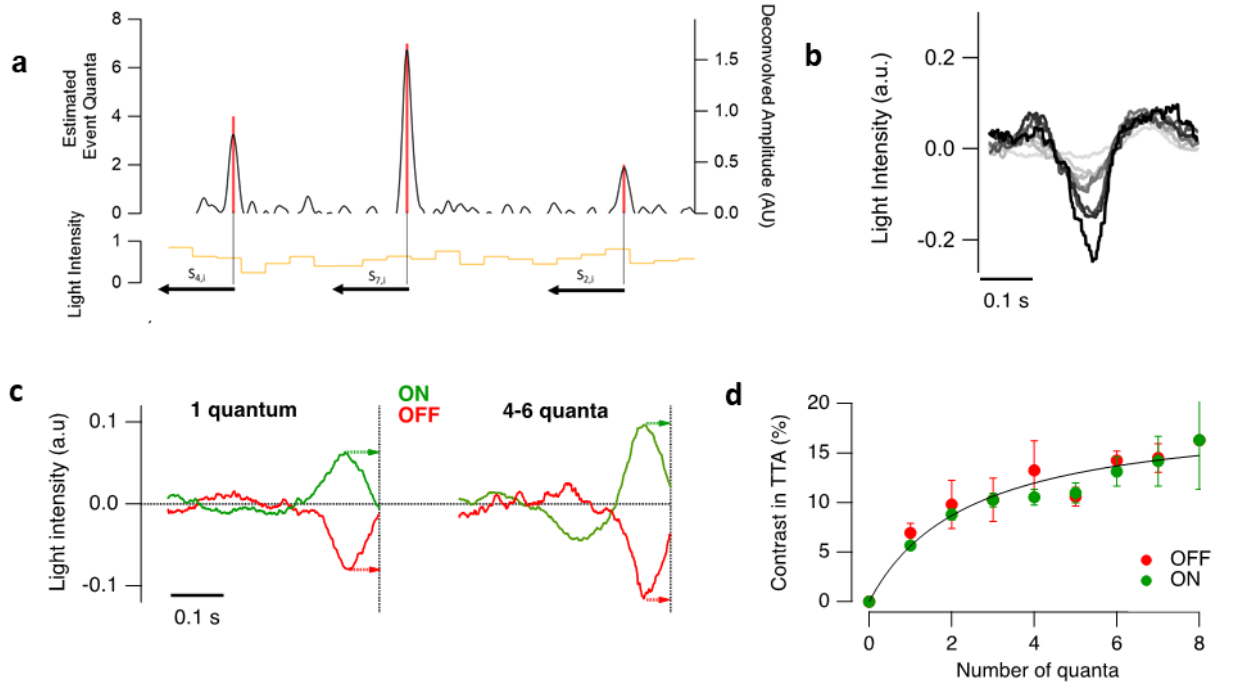

Figure 4.4: The Transmitter Triggered Averages show Contrast Dependence.

a) Example of the method of calculating the TTA. The stimuli preceding each quantal event type (top, red lines) are averaged (bottom, black arrows). **b)** Example of a single AZs TTAs. Here, the TTAs for events containing up to eight vesicles (darkness indicating increasing quantal content) are plotted. Note that the fewer quantal, lighter TTAs show less contrast than the higher quantal TTAs (darker lines). **c)** Average TTAs for one quantum and four to six quanta for ON (n=9) and OFF (n=8) cells. Note the increasing contrast for the higher TTAs. **d)** Plot of contrast in TTA as a function of quantal content – note the positive slope of the function. The relation could be described by a first-order saturation in the form $C = (C_{max})N / (N + N_{1/2})$, where $C_{max} = 19\%$ and $N_{1/2} = 2.4$.

To further analyse the contrast dependence in TTAs, we average the resulting filters for **N** cells, after splitting into ON and OFF. Examples for the uniquantal filters and 4-6 quanta filters are shown in **Fig4.4c**. For each average filter, we then computed the Michelson contrast in the filter (max-min)/(max+min) and plotted these data in **Fig4.4d**. As the number of vesicles in an event increases, so does the average contrast of the stimuli preceding the event – another strong indication that MVR is able to transmit information pertaining to visual contrast by utilizing an amplitude code. Note that for technological reasons, we were not able to compute the spatial TTA. Further experiments are required to investigate what spatial properties of a cells receptive field might vary with quantal content.

4.3.3. Higher quantal events convey more information per vesicle

In order to quantify how much information can be conveyed by each quantal event type, I computed the specific information ($I_2(S;q)$) between the distribution of stimuli S and each quantal event **q** observed, where the distribution **Q** is computed empirically by binning the response into 20 ms bins such that no two events occurred in the same bin and computing the proportion of bins containing each quantal event type. Like mutual information, specific information quantifies the degree of dependence between the stimulus and the response. However, mutual information quantifies the average information gained after observing all symbols (averaging over the probability of each symbol). Specific information, on the other hand, computes the amount of information gained from observing a *specific* symbol, without considering its probability. This relationship can be better understood by the fact that the dot product of the specific information vector and the probability of observing each symbol is equal to the mutual information. Note that for this analysis, the number of distinct responses possible is at maximum twelve – one for each observed quanta/bin from zero to eleven. Each contrast was delivered for a minimum of 20 s, ensuring that a minimum of 1000 bins are recorded for each stimulus. Consequentially, the bias – computed as the ratio of the number of observed responses and the number of distinct responses (Panzeri and Treves 1996, Panzeri, Senatore et al. 2007), is negligible.

As the distribution of quanta per event is contrast dependent, we find that the specific information is positive for all values. Thus, the presence of each quantal event type carries information about the stimulus to some extent, and no quantal event type is completely independent from the stimulus. Additionally, we find that as the quantal content increases up to five vesicles so does specific information. In order to quantify how efficiently these quantal event types are operating, we divided the specific information for each vesicle by the number of vesicles in the event (uni-quantal information is divided by one, 2-quantal information is divided by two, &c). Notably, this value also increases with quantal content, as can be seen in F**ig4.5a,b**. This indicates that higher quantal events are utilizing vesicles more efficiently than lower quantal events.

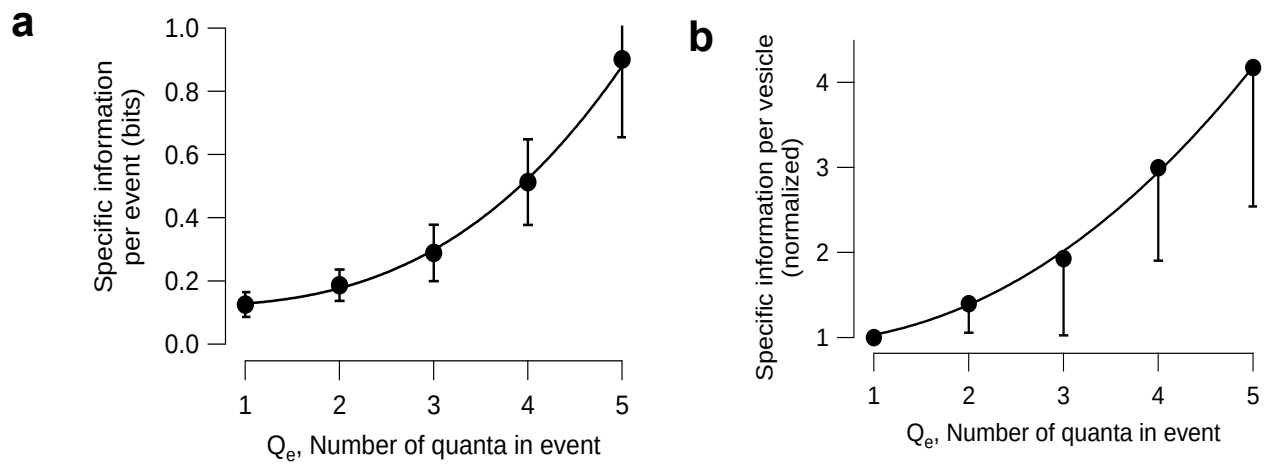

Figure 4.5: Higher quantal events convey more information than lower quantal events.

a) Specific information per event (I_2 bits) as a function of Q_e, the number of vesicles comprising the event. The curve describing the points is a power function of the form $i = i_0 + A \times Q_e^x$, with $i_0 = 0.12$ bits, $A = 0.008$ and $x = 2.8$. Results pooled from n = 17 synapses. **c**, Specific information per vesicle normalized to the value measured for a uniquantal event, and the curve describing the points is a power function of exponent 2.1. Results from the same 17 synapses in **b**. All data show mean ± s.e.m

Here, we note that although higher quantal events convey more information, they may not significantly affect the mutual information. While the do convey more information, they are less likely to occur.

4.4. Discussion

In this chapter I have demonstrated that BCs are capable of utilizing a system consisting of multiple synaptic symbols corresponding to the number of vesicles in an event in order to transmit information about a visual stimuli's contrast. While the phenomenon of MVR has been known for decades, little information exists on how the phenomena operates in intact systems. Most notably, there is a paucity of information regarding the functional implications of MVR in the retina. Does the tight temporal coordination of quantal release in MVR (Singer, Lassova et al. 2004) allow for a non-binary system of information processing? While previous work has addressed the issue of transmitting information using multiple symbols, such as is the case in many analyses on spike bursts (Carrillo-Medina and Latorre 2018, Zeldenrust, Chameau et al. 2018, Zeldenrust, Wadman et al. 2018), I here show that this framework also applies to synaptic

transmission. Even at the level of the synapse and timescales smaller than many RGCs membrane time constants, BCs can reliably transmit information by not only modulating the rate of synaptic events, but also the number of quanta released in a given event. This gives the first demonstration to my knowledge of synaptic coding using multiple synaptic signals. Rather than simply altering the rate of vesicle release in a standard Poisson fashion (Korn, Sur et al. 1994, Johnson 1996), BCs are capable of compressing multiple quanta into one release event, allowing for the formation of a new probabilistic symbol. Notably, we find that these 'higher quantal' symbols effectively transmit information regarding the visual stimulus.

While this work has focused on the how BCs respond to temporal contrast – specifically how a BC is capable of transmitting information about a stimulus's contrast by increasing the number of quanta in a synaptic event – it is likely that temporal contrast is not the only factor of the visual scene that can be encoded using an amplitude code. In fact, it is highly likely that amplitude coding will provide more of an impact on transmitting information regarding a stimulus's temporal frequency, due to the increased temporal precision of higher quantal events. Further experiments are required to verify these hypotheses, as well as to extend these analyses to more complex aspects of the natural environment, such as object movement and higher-order correlations.

4.4.1. Shifting from Binary

Here, we showed how the BC terminal can encode information not simply by altering the rate of univesicular events, but also by releasing multiple glutamatergic vesicles nearly simultaneously in MVR. In doing so, the nervous system is effectively increasing the number of distinct symbols with which it can operate – rather than a simply binary system, we now see a 'multi-nary' system, with symbols corresponding to the release of zero up to approximately ten vesicles. How might this shift from binary benefit the nervous system? The most obvious benefit is that by increasing the number of symbols in a system one can increase the entropy, and consequently, the mutual information. It is trivial to show that the maximum entropy distribution of a set of n symbols is uniform, where each symbol has equal probability, yielding an entropy of $log_2(n)$ bits. Speaking on a purely information theoretical level, then, it is simple that increasing then number of symbols will increase the entropy of the system. However, this does not take into account the quantitative nature of our symbols – each quantal event type by its very nature is associated with a cost in terms of its vesicular content. In the case of our system, a uniform distribution over n quantal event types (from zero to n-1 vesicle events) requires $\frac{n-1}{2}$ vesicles

per bin. Simply increasing the number of bins of a uniform set of quantal events would require a vast number of vesicles, as can be seen in **Fig4.6a**, where I plotted the entropy of a system of n symbols against the expected vesicles released in a second (assuming a bin size of 20 ms) for a uniform distribution consisting n symbols. Thus, the first point, corresponding to a binary zero or one vesicle event system, achieves its maximum entropy of $log_2(2)/.02 = 1/.02 = 50\ bits/second$ with the average release rate of $\frac{n-1}{2(.02)} = 25$ vesicles per second. Doubling the entropy requires three times the vesicles, and tripling it requires seven times the vesicles. Clearly, these values can quickly exceed reasonable levels for a single ribbon. A better question to ask, then, might be how can the nervous system switch to a multiple symbol system while constraining for the mean number of vesicles released? In fact, this is a known distribution, the Boltzmann distribution, given by probability mass function:

$$p(x_i) = \frac{1}{C} e^{-\varepsilon_i \beta}$$

with normalizing constant $C = \sum_{i=1}^{n} e^{-\varepsilon_i \beta}$. Originally constructed to calculate the probability of various thermal states in system as a function of the systems total energy, we can also find that this distribution is the maximum entropy distribution for a finite discrete set of symbols.

For example, assume we have a binary zero or one vesicle synapse. Binning the release events into 20 ms bins as previous, we can compute the entropy of the system as a function of its release rate, as in **Fig4.6b**. As expected, the entropy is maximized to 50 bits/second at a release rate of 25 vesicles per second, corresponding to a uniform distribution. Now, consider a different system operating with three symbols corresponding to zero, one, and two vesicle events. Then we can maximize the entropy of this system constraining the release rate, yielding a Boltzmann distribution, as seen in **Fig4.6c.** Notably, while this distribution peaks (at a uniform distribution for a three-symbol system), it occurs far later than the binary entropy peak. Additionally, the Boltzmann distribution always has at least as much, if not more entropy, than the binary system. In fact, one can *always* increase (or maintain) the entropy of a discrete system while maintaining a constant vesicle release by simply adding an additional symbol, although the advantage of adding more and more symbols quickly vanishes. Compare this to the 11 parameter Boltzmann distribution (corresponding to release events from zero to ten), as well as the geometric distribution – the maximum entropy distribution with constrained mean and infinite bins.

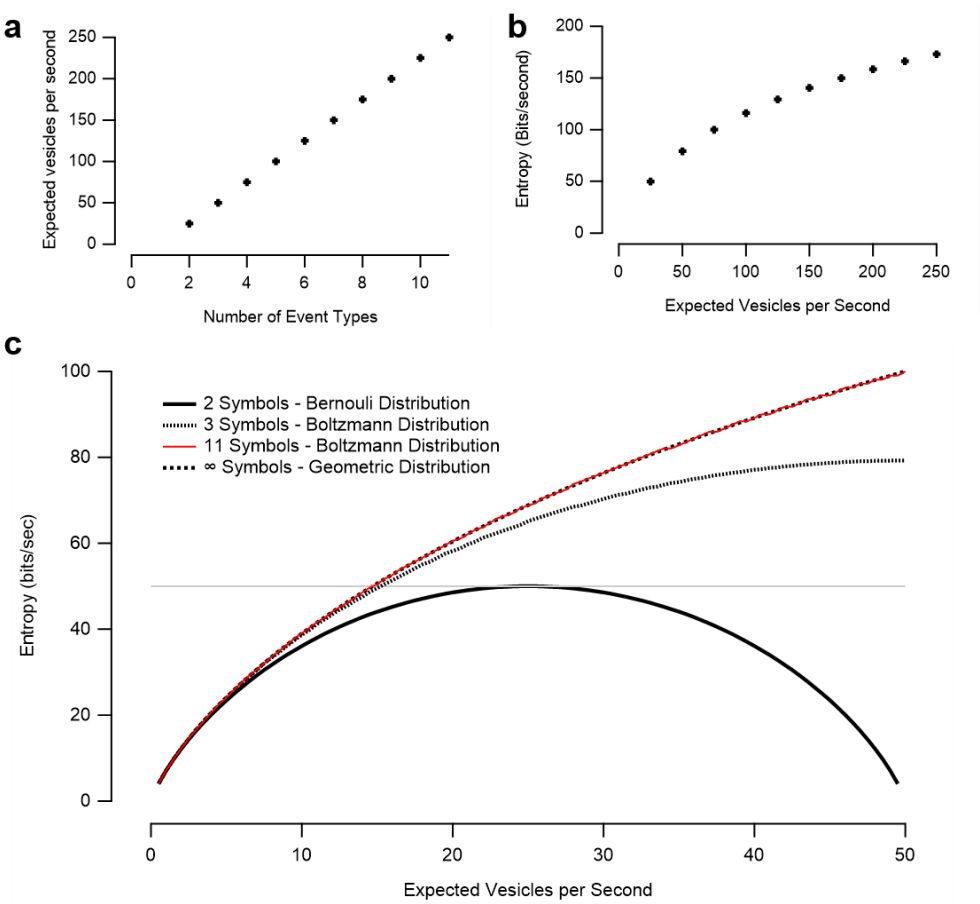

Figure 4.6: Differences in entropy and expected vesicles between Uniform and Boltzmann Distributions.

a) The expected number of released vesicles for maximum entropy uniform distribution consisting of from two to eleven signals. The number of vesicles linearly increases with number of symbols. Here, the bin size is assumed to be 20 ms**. b)** The entropy of a uniform distribution plotted against the expected vesicles released in a). **c)** The entropy of a Bernoulli distribution as a function of its expected vesicles released (black), compared with the entropy of a three-symbol system with the same mean release rate (fine dashes), given by a Boltzmann distribution. Increasing the number of symbols used can increase entropy while maintaining release rate. Compare this to the entropy of a distribution with the same mean but eleven symbols (coarse dash), and the entropy of a system with the same mean but infinite symbols – corresponding to a geometric distribution (red).

4.4.2. Utility, the Boltzmann Distribution, and Natural Statistics

In the previous section, we showed that one can increase entropy while constraining vesicle release by increasing the number of vesicles in an event, similar to MVR. One aspect of the distribution is that as the value of the symbol increases its probability decreases, as a

consequence of mean constraint. From a neuroscientific point of view, how can we imagine this distribution of vesicular release to benefit an organism? From a straight information theoretical standpoint, there is no quantitative differences between the quantal symbols – the ordering or value of the symbol is irrelevant, only their distribution. Thus, less-probably, surprising symbols, while they do convey a great deal of information, needn't have a subsequently larger effect on the post-synaptic cell. However, if we assume that increasing the number of vesicles in an event increases the probability of the signal being transmitted to downstream neurons, then we can naturally understand how this amplitude coding could benefit. Higher quantal events, being less frequent, can carry more information and induce larger activity in the cell. Thus, one efficient way a neural system could utilize MVR is by encoding rare and important (events with high utility) by adjusting the amplitude of the event. As these events are rare, they will not contribute much to the total vesicles released, but they are better able to allow for the signal to be reliably transmitted to downstream neurons. Consequently, it would be interesting to link the presence of an amplitude code and MVR with high-utility signals, such as predator avoidance or predation.

In the present work we focused our analysis on the responses of BCs to various contrasts to a full field amber stimulus, assuming that the distribution of contrasts was uniform. However, it is unlikely that the distribution of contrasts in a zebrafish larvae's natural environment would follow such a distribution. While recent work has begun to characterize the statistics of the zebrafish's natural environment as has been accomplished for terrestrial mammals such as the cat and mouse, the majority of these works have focused on the spectral differences in the zebrafish's environment, with a particular emphasis placed on color processing (Zimmermann, Nevala et al. 2018, Baden and Osorio 2019). Additionally, while the distribution of spatial frequencies in terrestrial natural scenes can be well approximated by a power law (Atick and Redlich 1992, Ruderman 1997), to my knowledge no such characterizations for the distribution of contrasts in underwater natural scenes exists. It would be interesting to see in further work how well the system is optimized to the organism's natural environment by recomputing the mutual information and specific information using the true distribution of contrasts in natural scenes, as has been shown in other organisms and other aspects of the visual environment (Zimmermann, Nevala et al. 2018, Baden, Euler et al. 2020, Zhou, Bear et al. 2020). If either metric were optimized with stimulus distributions matching those found in natural scenes, it would be likely that the system is highly adapted to the aquatic environment.

4.4.3. The TTA and Adaptation

Here, I showed the MVR is contrast dependent. One manner in which I did so was by using the fact that the contrast present in recovered TTA filters increases as a function of higher contrast. Following on the logic of the traditional STA, this would seem to indicate that higher quantal events have a dedicated higher contrast filter – that a specific quantum 'likes' a specific contrast preferentially. Indeed, this is most likely not the case – yes, the higher quantal events are more likely to be triggered by higher contrasts, but this ignores one of the ribbon synapses most profound properties – that of adaptation. Rather than each quantal event type being triggered selectively by a specific contrast, it is more likely the number of vesicles triggered by any contrast is highly dependent upon the number of available releasable vesicles. If we view the TTA in this sense, then it would perhaps be better to describe the contrast dependence not in terms of discrete number of vesicles released, but rather as some metric of the percentage of available vesicles. If ten vesicles are available for release, a 10-quantal filter would represent the stimulus capable of releasing all vesicles. If only 5 vesicles are available for release, then no 10-quantal filter is possible – the 5-quantal filter would now be the previous 10-quantal filter, representing the release of all available vesicles. While this subtlety might seem to increase the complexity of the problem, it ties very nicely into the adaptive properties of the ribbon synapse.

One manner of testing this would be to adapt the axon terminal to variable states before recording responses to white noise. If filters recovered after adapting the cell to high-contrast stimuli are shifted towards lower contrasts relative to unadapted extracted filters, it would indicate that the TTAs do not trigger the release of precise numbers of vesicles, but a proportion of the vesicles available for release.

4.4.3: Potential Mechanisms

Given the experimental evidence that individual vesicles are released from ribbon synapses in a coordinated and non-independent manner, what assumptions can we make about the mechanism generating release? Clearly, the standard, one-site/one-vesicle hypothesis does not apply – if, in fact, each potential site of vesicular docking can independently be released, the statistics of release would follow a binomial distribution and there would be no evidence of correlated release. A more scientifically plausible hypothesis for the mechanism of vesicle

release from ribbon synapses, then is compound fusion (Neef, Khimich et al. 2007), where multiple vesicles first fuse with one another before fusing with the cell membrane and being released. This presynaptic fusion would establish a dependence upon the release of individual vesicles, as is commonly found in experimental evidence. However, additional evidence suggests that this presynaptic fusion could result in delayed increases in glutamate concentration in the synaptic cleft relative to vesicles release in the traditional manner (Fuchs 2005). As some evidence suggest that the time course of glutamate concentrations is faster than this mechanism would predict, others have proposed alternative hypothesis for the mechanism of vesicle release. A ‘stamp’ mechanism, for example, would induce both correlated vesicle release as well as increased speed of glutamate concentration. Further experiments will be necessary to precisely uncover the responsible mechanism.

Chapter 5: Theoretical Frameworks for the Analysis of MVR

5.1. Introduction

A striking feature of ribbon-type synapses is their ability to release several vesicles within a period of a few milliseconds a process termed multivesicular release (MVR) (Singer, Lassova et al. 2004). In retinal bipolar cells and mechanosensitive hair cells the fusion of several vesicles can even be synchronized to within 100 µs (Mennerick and Matthews 1996, von Gersdorff, Sakaba et al. 1998). This process has been termed coordinated multivesicular release (CMVR) to highlight the deviation from a Poisson process in which vesicles fuse independently of each other. Nonetheless, multiple vesicles released in a time window of 100 µs or 10 ms will summate similarly on a postsynaptic ganglion cell because these have membrane time constants in the range of 10–40 ms, with as little as one ms in certain RGCs (O'Brien, Isayama et al. 2002). Glutamate release events comprised of different numbers of vesicles can therefore be considered as different synaptic symbols and recent work has shown that they can be utilized in an amplitude code, where information is transmitted by altering the distribution of event amplitudes in addition to the rate of events (James, Darnet et al. 2019).

Despite how common the presence of MVR in the nervous system seems to be, little information exists on how post-synaptic cells might encode synaptic information in the form of an amplitude-based code. How does postsynaptic cell machinery affect information encoded in both rate-based and amplitude-based inputs? In what situations would amplitude or rate inputs transmit the most information? How is this affected by postsynaptic cell physiology?

An experimental assessment of how MVR, even in early sensory systems, impacts on the synaptic transfer of information is extremely difficult because it requires recording spikes in the post-synaptic neuron while also characterizing the influence of individual synaptic inputs at which it is possible to count the numbers of vesicles released. The experimental reality is that a patch pipette placed on a neuron usually records a signal influenced by *all* synapses, although an important exception is the case of primary auditory afferents of Type 1, which receive inputs from individual ribbon synapses of hair cells (Glowatzki and Fuchs 2002, Fuchs and Glowatzki 2015). To make a more general investigation of the impact of MVR on the transmission of sensory signals we used a modelling approach in which the driving signal was a sinusoid mimicking the graded potential generated by a tone or flickering light (Pillow, Paninski et al. 2005, Schwartz, Pillow et al. 2006). This analogue input controlled the release of excitatory vesicles, either as a pure rate-code or as a hybrid rate-amplitude code, which then served as

the input to a conductance Leaky Integrate-and-Fire (LIF) model of the post-synaptic neuron (**Fig5.2a**) (Gerstner, Kreiter et al. 1997, Abbott 1999, Burkitt 2006, Burkitt 2006). The spike trains generated in the neuron were then analyzed using information theoretic metrics to assess how information transmission was affected by variables such as the amplitude of the miniature potential, the size of the post-synaptic neuron and the number of synaptic inputs it receives. The effects of either a rate-based or amplitude-based vesicle code on a post-synaptic neuron will obviously depend on the numbers of vesicles released so, to make a fair comparison of how effectively these two regimes might utilize synaptic vesicles to transmit information, the inputs to the model neuron were constructed such that the average number of vesicles released were equal at all times (**Fig5.2b**).

For this reason, we adopted a statistical and modelling approach. In order to answer these questions, we used a Leaky Integrate-and-Fire (LIF) model (Stein 1965, Burkitt 2006, Burkitt 2006) to simulate spiking activity of a neuron receiving stochastic input consisting of either a pure rate-based code or a hybrid rate-amplitude code. By analyzing the spiking output of the model using a variety of information theoretic metrics, we compare how efficiently synaptic vesicles are used to transmit information to a postsynaptic neuron. Modelling was based on experimentally measured features of MVR transmitting visual information from bipolar cells to retinal ganglion cells, but with the aim of understanding general properties of MVR in other contexts. How does MVR benefit the cell? Specifically, how are the statistical properties of the signal altered by the introduction of both an amplitude code (in the case of MVR in vesicular release) and a graded-to-binary switch (in the case of the ribbon-synapse to spiking cell?)

Amplitude coding and MVR enable more efficient information transmission in various situations. Faster membrane leak constants, and longer refractory periods increase the relative efficiency of amplitude coding over rate coding. However, rate coding is more effective with higher postsynaptic input resistance and a smaller vesicles/spike ratio. While amplitude coding suffers from a reduction in information contained in spike times, it is able to generate more spikes in postsynaptic cells with a given vesicular input, increasing the information contained in the spike count. MVR and amplitude coding thus can better guarantee that signals will be produced in postsynaptic neurons.

An example of the phenomenon of MVR, recorded during transmission of the visual signal from bipolar cells to retinal ganglion cells, is shown in **Fig5.1**. Here we used multiphoton

microscopy to image the glutamate reporter iGluSnFR (Marvin, Borghuis et al. 2013) in the retina of larval zebrafish, an approach that can count vesicles released from individual active zones (James, Darnet et al. 2019). The amplitude of release events varied widely during a sinusoidal modulation in light intensity (60% contrast, 5 Hz). The key observation is that the synapses transmitting the visual signal to the inner retina encode contrast as increases in both the average rate and amplitude of synaptic events (**Fig5.1b** and **c**). Although MVR has not been directly measured during responses to sound, both the rate and amplitude of synaptic events in primary afferents receiving inputs from auditory hair cells depend on the degree to which the presynaptic calcium current is activated, making it likely that a hybrid rate-amplitude code is also a feature of the first synapse in the sense of hearing and balance (Goutman and Glowatzki 2011, Li, Cho et al. 2014). Additionally, MVR has been observed in cells outside the sensory systems, such as hippocampus (Christie and Jahr 2006), cerebellum (Auger, Kondo et al. 1998), and somatosensory cortex (Huang, Bao et al. 2010). However, counting vesicles *in vivo* in central synapses is considerably more difficult – the tissue is deeper in the organism, making neurotransmitter imaging experiments difficult, and the MVR activity of these cells is often a direct consequence of spiking activity, requiring a disentangling of spike signal vs. neurotransmitter signal.

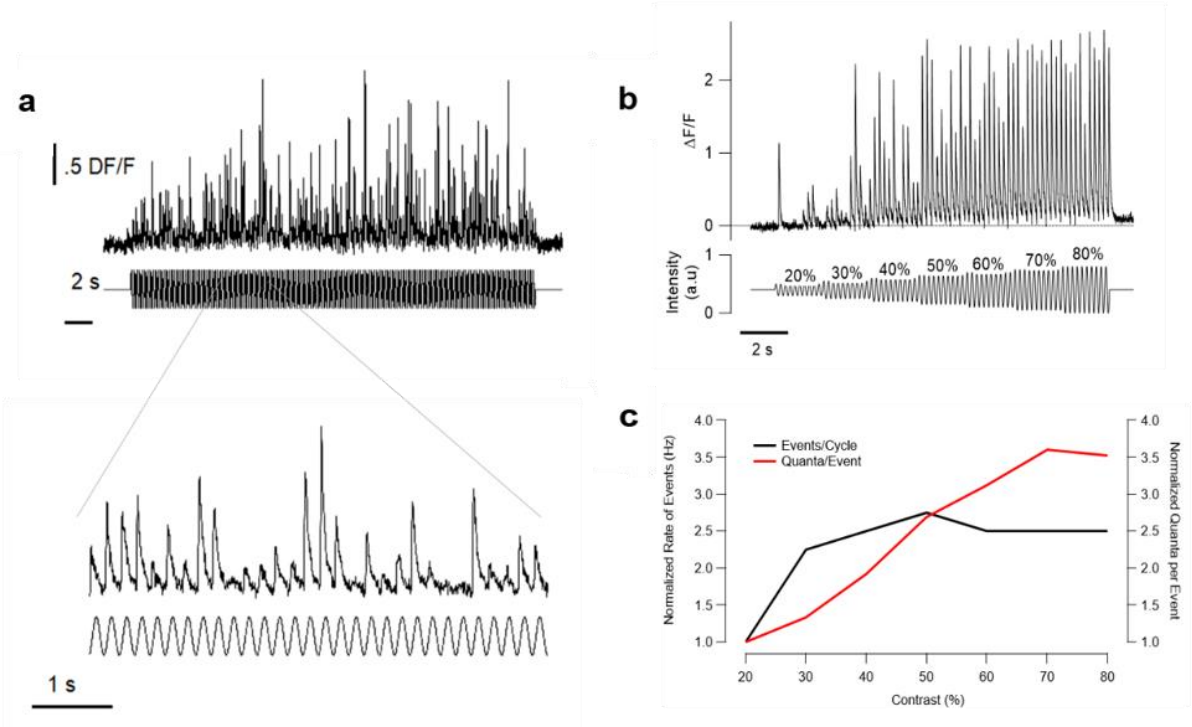

Figure 5.1: Information is encoded by modulating the amplitude of events.

a) Multiphoton imaging of glutamate release from a single zebrafish active zone in response to a five Hz full-field 60% contrast sinusoid. Note the large variations in amplitude of glutamatergic events, corresponding to multivesicular release events. A blowup of the above trace is presented below. **b)** Glutamate release in response to 5 Hz sin waves of increasing contrast. **c)** Normalized change in the rate of events (black) and average quantal content (red). Note that this synapse reaches a maximum event rate at a lower contrast than it reaches the maximum quantal rate, indicating that information beyond simple rates can be encoded by modulating the amplitude of glutamatergic events.

5.2: Methods

A basic aspect of our approach was to use Poisson Processes (PPs) to describe the dynamics of the synaptic outputs. Ignoring mathematical rigor, a PP is a mathematical manner in which to describe the random placement of points in a dimension (such as space or time). Poisson Processes have been used extensively in neuroscience as a simple way in which to describe the location of neural events in time, most commonly action potentials (Gerstein and Clark 1964, Moore, Perkel et al. 1966), mathematically simple enough to easily work with, but carrying many interesting properties that can match well with the properties of binary signals. For example, the unitary relationship between spike count means and variance matches those of certain spiking cells, and the memoryless property can be useful for early sensory cells.

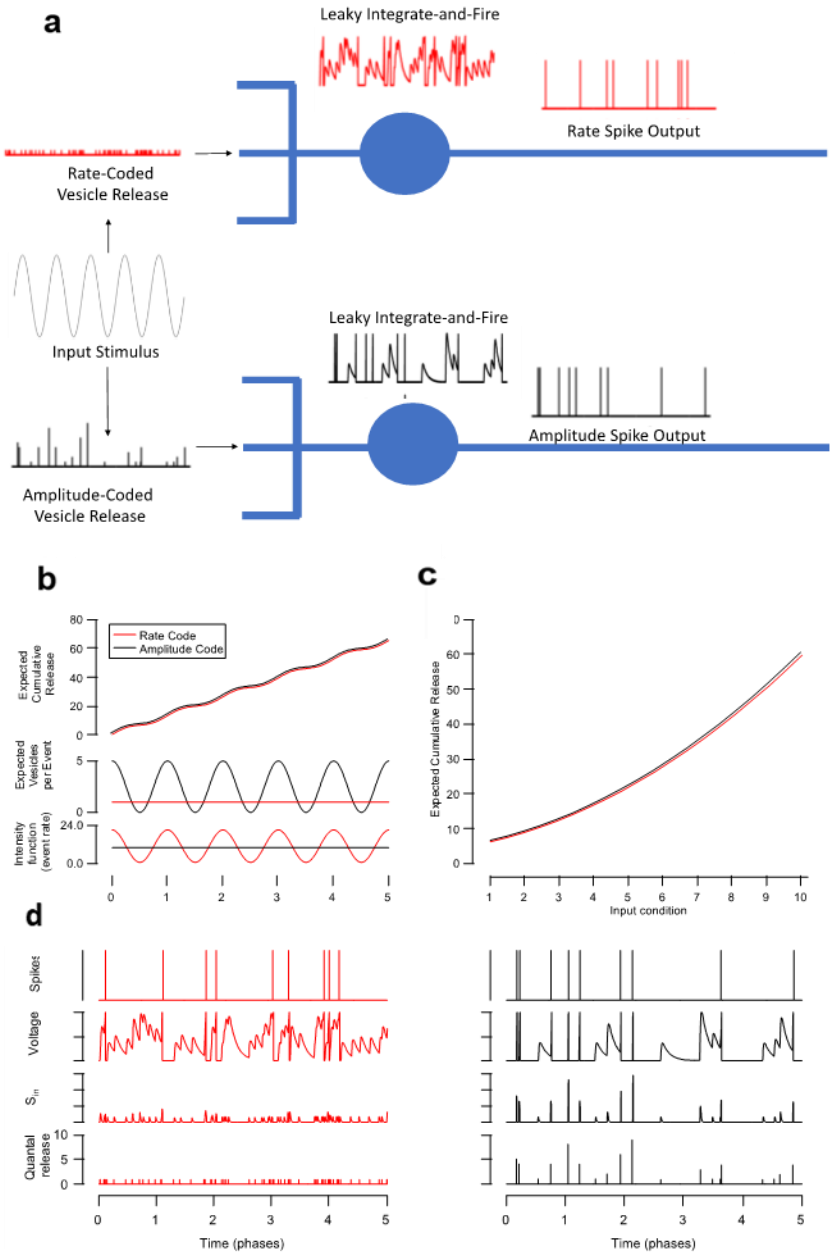

Figure 5.2: Illustration of technique.

a) A time series input (gray) to a cell is encoded by either modulating the rate (top, red) or amplitude (bottom, black) of vesicular events. The inputs are passed through a Leaky Integrate-and-Fire Model, and the spiking output of the modelled cells are compared. Amplitude and Rate modulation of vesicle release. **b)** bottom: intensity function driving the rate of vesicular events. Middle: expected number of vesicles released in an event as a function of time. Top)

Expected cumulative vesicles released as a function of time for rate and amplitude codes, offset for visualization. These functions have been equalized by design. **c).** Expected total vesicles released over a one second window as a function of stimulus 'contrast'. **d)** Example time series for a synaptic output, the induced current, voltage of a postsynaptic cell, and spiking output for a rate and amplitude coded cell.

The simplest form of PP is the Homogeneous Poisson Process, often implicitly used to model spontaneous neural activity. Here, the probability of an event occurring in any given time is a constant λ. Nonhomogeneous Poisson Processes (NHPPs), in which the probability of an event varies in time according to a function $\lambda(t)$, are used to model more complex temporal dynamics of neural activity (Johnson 1996). We used such a NHPP as the input for the rate case, in which the mean rate λ varied sinusoidally in time (**Fig5.2a** and **b**; see Materials and Methods). By definition, simple Poisson Processes describe a situation in which no two events can occur simultaneously, and therefore, cannot be used to model MVR when two or more vesicles are released within a single synaptic event, as we define here. Mathematically put using little-o notation, the probability of two or more events occurring in a time window T is o(T), where o(T) is defined such that $\lim_{T\to 0} \frac{o(T)}{T} = 0$. To construct an amplitude-modulated Poisson Process we turned to a modification known as Poisson Splitting or Location-dependent thinning (Resnick 1992), where events can be assigned a type based upon their location in time (see Chp 2 for more information). In this case, events occur with a constant rate, and the amplitude (or number of vesicles in an event) for each event is chosen based upon when the event occurred (**Fig5.2b**). We can thus use drive the LIF model neuron in three distinct regimes using defined statistics: a pure rate case, where all information is encoded by modulating the mean rate of release of individual vesicles; an amplitude case, where all information is encoded by modulating the amplitude of vesicular events occurring at a constant rate, and a hybrid case that combines aspects of both rate and amplitude coding. For simplicity, we used a sinusoidal intensity function for the rate case, and a Binomial compounding function with a constant N=10 and where the probability of release is a sinusoid resulting in equivalent release in the amplitude case

Before we begin to discuss the results of the models, it is useful to consider a few facts that might not be overtly obvious. Firstly, while the vesicle release presented here takes inspiration from responses to temporal contrast in the early sensory systems, we do not explicitly model the function mapping from stimulus contrast to vesicle release. Rather, we assume the stimuli inducing spiking in postsynaptic cells arises not from the distribution of

sensory contrasts, but rather ‘input contrasts’. This was done largely to reduce model complexity – rather than first assigning a distribution of sensory contrast, using this contrast to compute vesicle release, and then analyzing spiking activity induced from this release, we simply assumed the input vesicle release series was the underlying information encoder and the spike the decoder. Thus, the independent variable in this situation is not the visual or auditory contrast of a stimulus, but rather the contrast of the function generating vesicle release. At the lowest release rates, inputs are in the form of a homogeneous process with expected release rate of five vesicles/sec. As release rates increase, so does the temporal structure of the signal, increasing the non-homogeneity in release (realized by increasing the amplitude of the sinusoidal generating release. While these amplitudes can be increased to arbitrary levels, we chose a maximum of 60 vesicles/second – corresponding to a release intensity contrast of approximately 80% contrast and matching release rates from single ribbons of BCs, where the rates are sinusoidal with a minimum of 5 ves/sec and a maximum of 60 ves/sec. The second property necessary to mention is the number of inputs. As Poisson Processes are summable, the model is agnostic to how many inputs a cell receives – the inputs are identical whether it is one cell with a set release function, or two cells each with half the set release function.

The impact of a synapse depends on the morphological and electrical properties of the post-synaptic neuron. For instance, RGCs excited by bipolar cells vary widely in the size of their dendritic trees and the membrane time-constants: in the fovea midget ganglion cells have small receptive fields and are electrically compact so a single vesicle is likely to have greater influence of the probability of a spike compared to an RGC in the peripheral retina, with a large dendritic tree and longer membrane time-constants. The parameters of the LIF model were therefore varied to investigate whether MVR had different effects on different types of post-synaptic neuron. We investigated the impact of these properties of the post-synaptic neuron in three basic ways. First, we held the number of synaptic inputs constant and varied the size of the excitatory conductance activated by glutamate released from a single vesicle, reflecting variations in the size of the miniature post-synaptic potential. Second, we varied the time-constant and membrane resistance of the post-synaptic target, reflecting neurons of different size and resting properties. Third, we varied the number of inputs while fixing the number of vesicles required to generate a spike, to investigate how the relative contributions of amplitude and rate coding might depend on the degree of convergence.

5.3: Results

5.3.1. MVR increases the efficiency of spike generation

We constructed the LIF model in such a way that *k* vesicles released simultaneously will generate just enough depolarization for the cell to spike from rest. In other words, *k* is the minimum number of vesicles released simultaneously that can trigger a spike (or the number of vesicles required to generate a spike in a perfect integrator). Values of *k* measured experimentally vary widely, both between and across cell types. While most ganglion cells require ten or more vesicles (Freed 2005) to spike, other RGCS, due to inherent passive electrical properties, can spike to even fewer vesicles (O'Brien, Isayama et al. 2002). For $k > 1$, we notice two main effects it immediately becomes apparent that amplitude coding is better able to generate spikes than rate coding over the one-second period simulated. In **Fig5.3**, we analyzed the spike count distribution for a one second period for the case of $k = 5$ at low contrast (**Fig5.3a**) and high contrast (**Fig5.3B**), comparing a pure rate-code (red) with a hybrid rate-amplitude code (black) with the same mean release rate. We also compare a small and compact target neuron (leak time-constant $\tau = 1$ ms) with a relatively large neuron ($\tau = 50$ ms). The second notable difference between the conditions is the reduction in variability and Fano Factor in the rate input relative to the amplitude input.

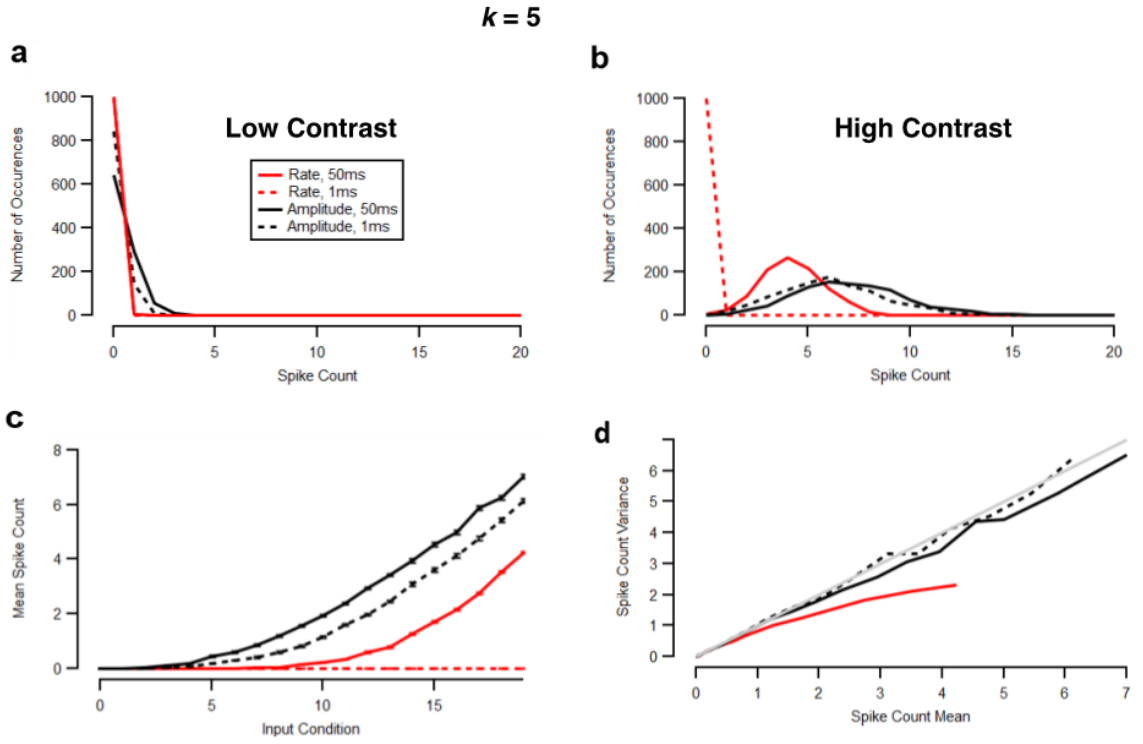

Figure 5.3: The more vesicles required to spike, the better amplitude coding is at generating spikes, and the leak time constant becomes more important.

a) Spike count distribution for low and **b)** high contrasts for when five vesicles are required to generate a spike. Note that rate coded inputs now generate fewer spikes than amplitude coded inputs. **c)** Mean spikes as a function of input. Note that the output is now dependent upon the leak constant, and a leak of 1 ms prevents all spiking activity in the rate coded input**. d)** Spike Count variance vs. mean. While the amplitude input is approximately Poisson, the rate input shows sub-Poisson variability.

The most striking effect of MVR was to increase the average spike count in the post-synaptic target neuron compared to the pure rate-code, and this effect was evident across the range of contrasts and for leak time-constants between 1 and 50 ms (**Fig5.3c**). In order to understand why this is the case, note that the value of τ determines the number of successive vesicle release events required to generate a spike within a given time interval. Consider a perfect integrator, where effectively $\tau = \infty$. There is zero voltage lost between input events, so the cell will almost surely (that is, with probability one) spike in response to any Poisson input where $\lambda > 0$. Introducing a finite leak time constant alters the spiking behavior – the cell will only spike if the rate of glutamatergic events sufficiently overpowers the leak. In the more realistic scenario of the leaky integrator, there is a nonzero probability that the cell will NOT spike in response to Poisson input, and this probability is controlled by the corresponding rate of Poisson input and time constant. For slower leaks, threshold for a spike can be achieved at a slower rate and still generate a spike. The lower the value of τ, or the higher the value of k, the higher the mean rate of vesicle release required to reach threshold and trigger a spike for a pure rate code.

In contrast, consider a cell receiving MVR input. Here, *any* MVR event composed of at least *k* vesicles will generate a spike, regardless of the value of the time constant. As the vesicles are released simultaneously, no voltage can decay from the cell, so the cell will spike with probability one. Then any MVR event composed of at least k vesicles can guarantee a spike – subthreshold dynamics only apply to events consisting of fewer than k vesicles. Phrased differently, with a finite leak constant, a purely rate input has a nonzero probability of never spiking. An amplitude coded input consisting of MVR events of k or more vesicles, on the other hand, is *guaranteed* to spike – at the very least in response to all MVR events consisting of up to k vesicles that it receives. Consequently, while the leak time constant τ affects both rate and amplitude coded inputs, it degrades spike generation through rate coded inputs to a higher degree. These simulations demonstrate that hybrid rate and amplitude coding have the potential to generate spikes more efficiently than a pure rate code in neurons with a range of sizes and input resistances and over a range of stimulus strengths.

5.3.4: MVR Ignores Convolutional Statistics

A second notable difference between rate and amplitude codes was the variance in spike counts, which is shown in **Fig5.3d** as a function of the mean spike count. Note that for increasing k the variance was approximately proportional to the mean when inputs to the neuron employed MVR, while a pure rate code demonstrated a Fano factor less than one. This variability in neuronal responses is commonly found in the nervous systems. Cells of the early sensory system, which require precise timing information, are capable of sub-Poisson variability. Cortical cells, which are more heavily influenced by common noise sources, on the other hand, show higher, super-Poisson variability (Shadlen and Newsome 1998, Goris, Movshon et al. 2014, Charles, Park et al. 2018). While some of these effects are due in part to the mechanics of the leaky integrate and fire model, more are due to the input statistics (Chatfield and Goodhardt 1973, Maimon and Assad 2009). The time before k events in a Poisson Process is not exponentially distributed, but rather Erlang distributed, with density given by convolution of exponential distributions:

$$f(t) = e_1(t) * \ldots * e_n(t) = \frac{\lambda^n x^{n-1} e^{-\lambda x}}{(n-1)!}$$

where e(t) are iid exponential distributions, n is the total number of exponentially arriving events before a spike, $*$ is the convolution operator, and f(t) is the Erlang probability density with mean $\frac{k}{\lambda}$ and variance $\frac{k}{\lambda^2}$. While the Erlang process models interspike times, and not distributions of spike counts, it is useful to understand the phenomena from this standpoint as the mathematics result in simple closed-form solutions – simply note that any decreases or increases in variability in interspike times result in a corresponding decrease or increase in spike count variability.

To begin, consider three perfect integrator cells. The first cell receives Poisson input with unitary constant rate and will spike in response to a single vesicle. Here, the time between spikes is, like the input, exponentially distributed, and the spike count variability is Poisson with variability one. The second cell requires two vesicles to spike but receives input at twice the rate as the first cell. The distribution of interspike times is no longer exponentially distributed like the inputs, but rather Erlang distributed with variability 0.5 and mean 1. This reduction in variability is then carried over to the distribution of spike counts, resulting in a Fano Factor below one. Consequently, any increase in k inherently reduces the variability, and increases the SNR, for the pure rate driven case – compare this to **Fig5.3c** red, for the rate input case which shows a sub-Poisson spike count variability. Consider now, like the previous example requiring two

vesicles to spike. However, rather than receiving input in the form of single vesicles, this cell receives unitary Poisson input in the form of two-quantal MVR events. As such, while a single *vesicle* does not generate the current required to depolarize the cell to spiking threshold, a single *event* does, and the spiking output follows a Poisson distribution. The spiking output, like the first cell, then has a unitary variance and Fano Factor, see for example the unitary relationship between spike mean and variance for the amplitude code inputs in **Fig5.3c**.

In this toy example, we can begin to understand how MVR can be a disadvantage to a cell. By grouping multiple vesicles into MVR events, the postsynaptic cell no longer benefits from the variance reduction inherent in convolutional statistics. The rate case, which allows for no MVR, is completely susceptible to this effect, but the amplitude case is less affected; it does not necessarily require k events to generate a spike in the amplitude case, but rather k vesicles. As multiple vesicles can be released simultaneously in the amplitude input, this reduction in variability is mitigated, as we indeed find (**Fig5.3d**).

5.3.4: The single-vesicle single-spike case

Varying the physical properties of the postsynaptic cell drastically alters a neurons response properties. Midget ganglion cells, which can receive input from a single BC, tend to spike to fewer vesicles. Additionally, although not found in the retina, recent work indicates that single vesicles released from an inner hair cell can reliably trigger a spike in the post-synaptic afferent (Rutherford, Chapochnikov et al. 2012, Grabner and Moser 2018). These fibers have an afferent boutons with input resistances on the scales of a GΩ, allowing a single vesicle to generate sufficient depolarization to generate a spike (Glowatzki and Fuchs 2002). We therefore also explored the behavior of the model in the particular case of $k = 1$ in order to better understand how MVR might play a role in information transmission between high input resistance cells. Note, that here we utilized the empirical ‘plug-in’ estimates for mutual information. Biases were avoided by utilizing 1000 samples for each estimate (see chapter 2).

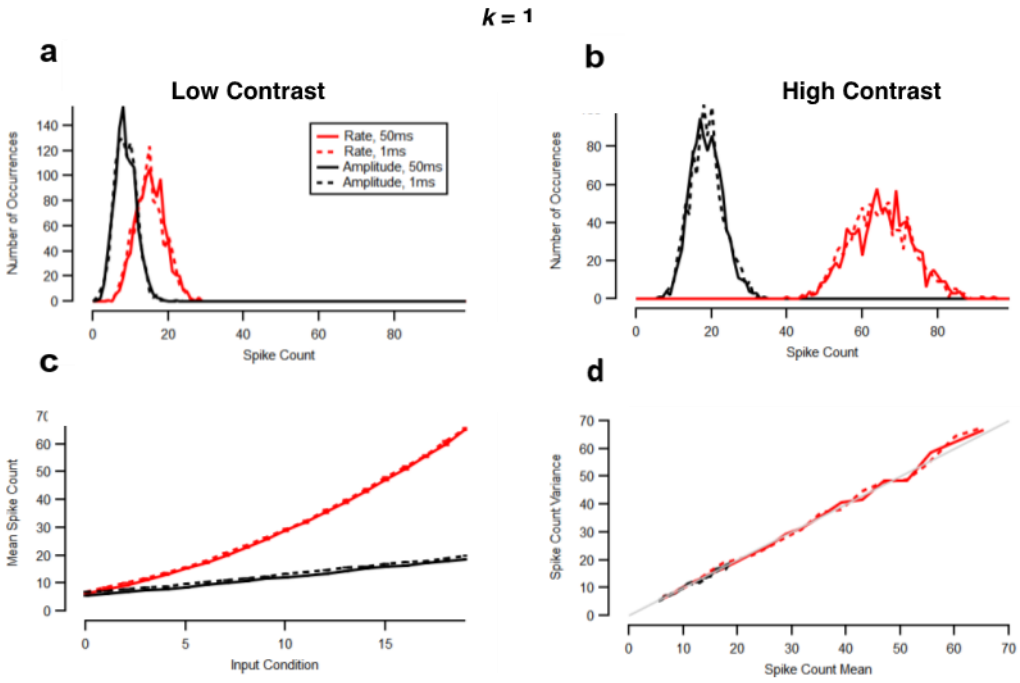

Figure 5.4: When a single vesicle is sufficient to generate a spike, amplitude coding reduces mean spike count, and the output spike is identical to the distribution of input events (a Poisson Process).

a) Spike Count distribution for low contrasts and **b)** high contrasts for cases when a single vesicle elicits a spike. Note that leak does not affect the distributions **c)** Mean spikes as a function of input contrast (+/- SE), offset for visibility **d)** Spike Count Variance vs Mean. Note that all conditions lie along the unity line (gray), indicating the outputs are still a Poisson Process.

MVR reduced the average rate of spike generation compared to a pure rate code, across the range of temporal contrasts (**Fig5.4c**) by as much as seven-fold at higher contrasts. To understand this, consider an idealized neuron receiving stochastic vesicle input encoding the contrast of a particular, discrete stimulus. In this scenario, the contrast of the stimulus is represented as a linear increase in the rate of vesicles released according to Poisson Process with rate $\lambda(t)$. Accordingly, the number of vesicles arriving at the neuron in any window (0,T) is Poisson distributed, with equal mean and variance. When a vesicle reaches the neuron, it generates a total depolarization of X mV, an amount sufficient to generate a spike from resting voltage. The voltage is then reset, and any left over current from the event is negated.

In this case, it is simple to see that the neuron will reproduce the vesicular input with perfect fidelity; every incoming vesicle produces an outgoing spike, and the statistics of vesicles

arriving to the neuron and the spike output are identical (ignoring any effects of refractory period). Thus, as the rate of vesicles increases, the number of spikes correspondingly increases, as well as the variability in the spike count. An analysis of the spike output of the model is shown in **Fig5.4** (red), where we plotted the distribution of spike counts in response to stimuli of increasing contrast for a cell with a lower rate of input vesicles (a), and a higher rate of input vesicles (b). Here, the increase in released vesicles during a simulated increase in visual contrast correspondingly increases spike output, and this relationship is plotted as a contrast response function in **Fig5.4c** (red). As is expected, the spiking output also follows a Poisson Process, here verified by the linear relationship between spike output mean and variance (**Fig5.4d**).

Let us now consider an alternative *hybrid* input, in which information is encoded by varying both the basal rate of vesicular events according to a Poisson Process, as well as the number of vesicles which compose them. As such, a single vesicular event can consist of multiple vesicles, mimicking the phenomenon of MVR. Note that the previous, rate-coded case can be considered a subset of the hybrid case where all vesicular events contain a single vesicle (thus, rather than having a stochastic function to dictate the number of vesicles in an event, all events deterministically contain one vesicle). Continuing with the above construction that a single vesicle is able to generate a spike, how would the spiking output of the neuron differ from the previous, rate-coded case, given that the same number of vesicles are used in both conditions? As a single vesicle generates enough voltage change to elicit a spike, the combining of multiple vesicles into a single event effectively wastes vesicles. The spiking output of the neuron will no longer follow the number of vesicles arriving, but rather the number of *events* that arrive. A decrease in the number of total events is necessary to allow for placing multiple vesicles in the same event if the total number of vesicles released is to be the same. If you reduce the number of events such that there are fewer events than there are vesicles released, at least one event will have more than one vesicle, and the overall spike output of the model is decreased relative to the rate-coded case. Thus, for this single-vesicle case, the distribution of spike counts for this hybrid input will be shifted towards lower values relative to those of the rate input (**Fig5.4a** and **b**, black). The resulting contrast response function has a shallower slope than the rate case (**Fig5.4c**), though the output is still Poisson (**Fig5.4d**).

Note that in our idealized neuron we have neglected any mention of voltage leak. In fact, in our scenario the spiking output is completely independent of the cell membrane time constant τ – as all events generate a spike and reset the voltage back to its resting potential, there are no

instances of sub-threshold depolarizations; there is no voltage to decay. To illustrate this, we included in **Fig5.4** samples from simulations with either 1 or 50 ms membrane constants.

5.3.5: Temporal Properties of the Spike Output

Up to this point we have considered a simple scenario in which synaptic events occur at a constant rate for a given stimulus contrast so that little information is contained in the times at which events occur. However, this is not generally the case – many cells encode temporal contrast by continuously varying the mean rate of vesicle release in time. A key question then is the temporal precision with which modulations in intensity can be signaled. In retinal bipolar cells and auditory hair cells larger MVR events have been shown to be more precisely timed relative to the stimulus compared to those comprising of fewer quanta (James, Darnet et al. 2019) (Li, Cho et al. 2014). Increasing temporal contrast therefore increases not only the rate of events and the distribution of amplitudes, but also the precision with which events occur in time.

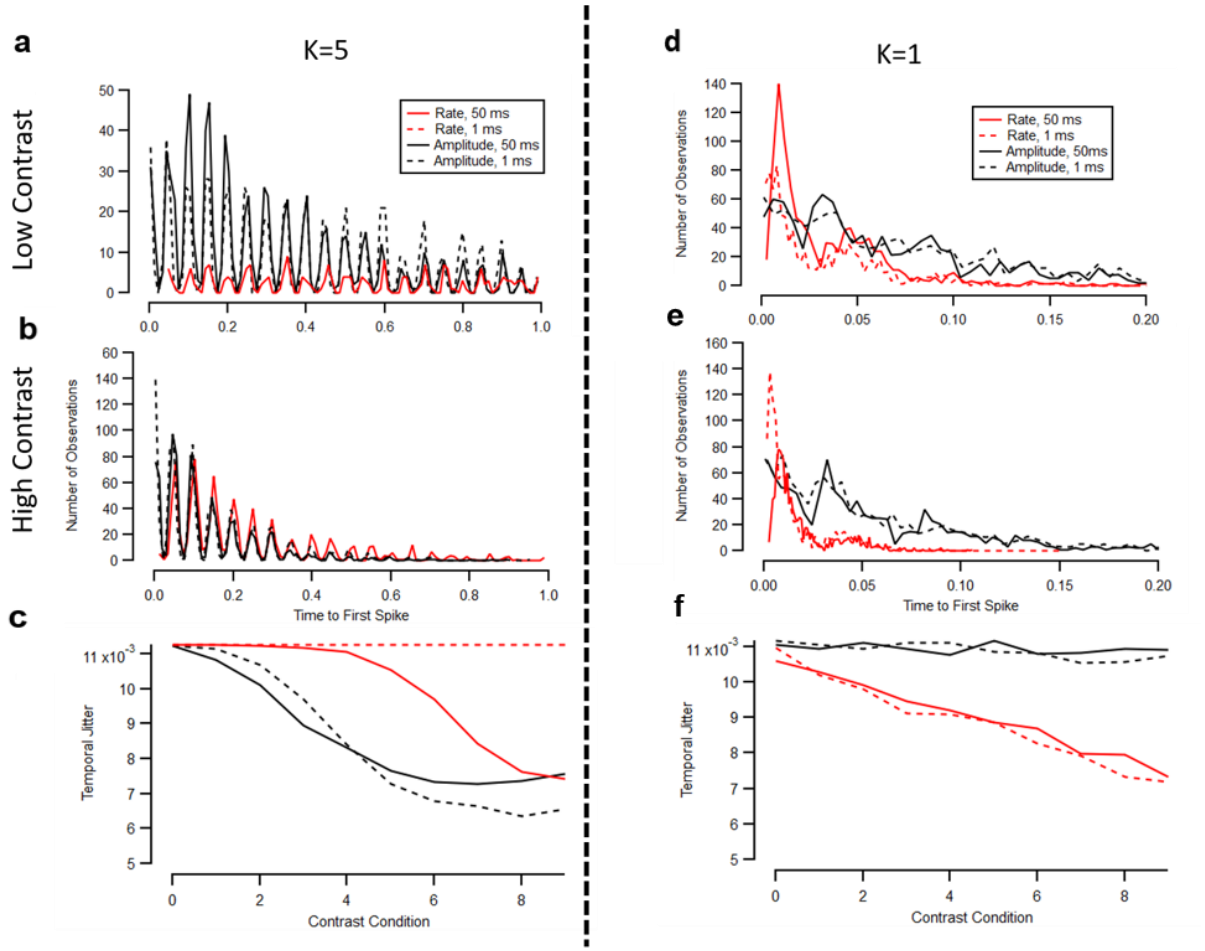

Figure 5.5: Benefit of amplitude and rate coding to spike timing precision depends upon postsynaptic parameters. A-C: Precision of spike output when a single vesicle generates a spike.

a) Distribution of first spike latencies for low and **b)** high contrast. **c)** The temporal jitter of the spike output as a function of contrast. Note that for the single vesicle case, the rate code consistently is more precise. **d-f)** Same as above, for the k = 5 vesicle case. Note that here, amplitude coding conveys higher spike precision than rate coding.

To investigate the impact of MVR on the precision of spike timing, we again used a periodic input while imposing the constraint of equalizing average vesicle release rates with the pure rate code. Within our model, higher quantal events are equipped with a higher temporal precision, as they are only elicited at the peaks of the sinusoids, while lower quantal events occur in a more dispersed fashion throughout the lower amplitudes of the sine wave. As previous, the release rates ranged from five vesicles per second at the lowest inputs levels, and 60 vesicles per second at higher contrasts.

Note that here the lowest contrast corresponds to a zero-amplitude sin wave with a baseline offset to mimic spontaneous activity. A direct consequence of this constraint is a reduced number of events in the hybrid code compared to the pure rate code. **Figure 5.5a** and **b** show a histogram of first spike times for the amplitude and rate case when k = 5 using a stimulus modulated at 20 Hz. Note that the temporal frequency was increased for this analysis in order to reduce the number of possible responses to a cycle by reducing their length relative to the previously used 5 Hz stimulus. A hybrid code generates more spikes than the rate code at low contrasts (**Fig5.5a**) but this advantage is lost at higher contrasts (**Fig5.5b** and **c**). Precision was then quantified as temporal jitter, the standard deviation of the phase of spikes. Over a range of contrasts, MVR improves temporal precision of spikes generation in both small, tight cells and larger leakier cells (broken and solid black lines in **Fig5.5c**). The pure rate code matches the temporal precision of the hybrid code only for the largest modulations in the signal driving synapses transmitting to a neuron with a long time constant.

5.3.6: Towards Information

The impact of MVR on spike generation (**Figs 5.3** and **5.4**) suggests that it might also affect information transmission across the synapse. We investigated this possibility by delivering stimuli of different contrasts to the LIF neuron and calculating the mutual information in the resulting spike responses (see Chp. 2).

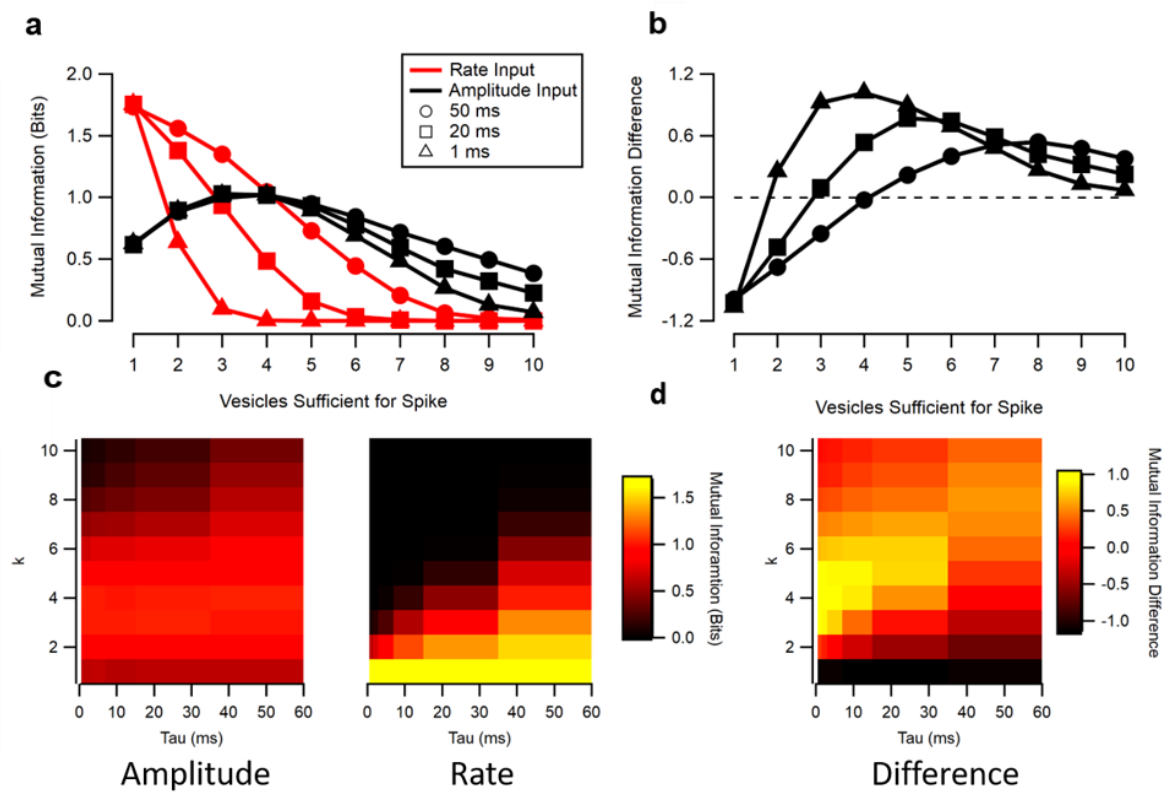

Figure 5.6: Spike Count Information for Amplitude (black) and Rate (red) inputs as a function of tau and k.

a) Mutual information for the rate input (dash) and amplitude input (solid) as a function of the vesicles sufficient to generate a spike. Values shown for tau of 50, 20, 10, and 1 ms. **b)** Mutual information difference (amplitude – rate) for the data shown in a. **c)** Mutual information (as a) in matrix form. **d)** Mutual information difference in matrix form.

To quantify these effects we computed the mutual information I(N;S) between the distribution of spike counts N and the distribution of contrast stimuli S for both the rate and the hybrid input cases. **Fig5.6a** plots I as a function of the relative spike threshold k for the case of a pure rate-coded synapse (red) and one employing a hybrid rate-amplitude code (black). The impact of the membrane time constant τ of the target neuron can be appreciated by comparing the plots for values of 1 ms, 20 ms and 50 ms. For the single-spike single-vesicle case, the mutual information is higher for the rate code and constant across values of τ, an expected outcome given that we have already observed that the distribution of spike counts is independent of tau (**Fig5.6**). As k increases, the amount of information a pure rate code can transmit decreases, largely as a result of the decrease in spike count. A matrix of the

interaction between these parameters is shown in **Fig5.6c** (left). Again, we see a profound effect of the membrane time constant, where faster leaks result in less mutual information than slower leaks (**Fig5.6a**). The general picture is that information transmission through pure rate-coded synapses is strongly dependent on the size and resistance of the post-synaptic neuron and the minimum number of vesicles required to generate a spike.

The involvement of MVR altered the transmission of information in several ways. First, the maximum transmission was achieved around k = 3-5, across a wide range of values of τ (black traces in **Fig5.6a**). Second, the decrease in I as *k* increased further was more gradual than in the case of a pure rate code input and also less dependent on τ, as summarized in **Fig5.6c** (right). To investigate compare these coding strategies we computed the information difference (Hybrid-Rate) for all conditions and plotted these as a function of k and τ in **Fig6b**. Across all values of τ, rate coding is able to transmit more information than amplitude coding when a single vesicle is sufficient to generate a spike. However, as k increases, we begin to see more information transmitted via hybrid coding than rate coding. For slower leaks, the peak I is achieved in neurons with longer time-constant i.e., requires integrating signals from multiple vesicles over a longer-time scale, allowing for spike generation to continue even when k is increased. Hybrid coding is therefore advantageous over a range of conditions, as shown by the diagonal streak in the information difference plot in **Fig6d**).

5.3.7: Spike Sequence Information

The analysis presented above relies only upon the distribution of spike counts in response to a stimulus; the precise timing of the spikes does not impact the calculation. To further study the possible effects of hybrid coding on spike timing mutual information, we simulated responses to a 20 Hz sine wave of varying amplitude. Each response was then discretized by binning the spikes into 50 1 ms bins such that no bin can contain more than one spike. Counting the number of occurrences of each distinct ‘word’ then yields a response distribution that takes support over all binary sequences of length 50. Mutual information can then be computed by noting how this distribution of responses shifts with varying stimuli, with the same logic as the simple spike count case. Note that here, we simulated 1000 1 s responses of 20 cycles, resulting in a total of 20,000 simulations for the mutual information estimate. While there is a possible 2^50 responses, we found less than 100 distinct responses, resulting in a negligible bias as seen by the ratio of response repetitions (Panzeri, Senatore et al. 2007). The switch from a 5 Hz stimulus to a 20 Hz stimulus was used to increase responses repetitions, and thus avoid any potential bias.

We first note that the information in spike sequences for either input condition is 0.8 out of a maximum possible 3.32 bits, as seen in **Fig5.7**. However, note that a different stimulus was used for the spike sequence information estimate – one cycle response of a 20 Hz stimulus as opposed to five cycle responses for the spike count information. As with the spike count information, we see for the rate input case a monotonic decrease in mutual information as a function of k, as well as a monotonic increase in mutual information as a function of tau. However, the falloff of mutual information in the spike timing case is faster than that of the spike count case, indicating that increasing the number of vesicles sufficient to generate a spike more radically affects information contained in spike sequences over spike counts.

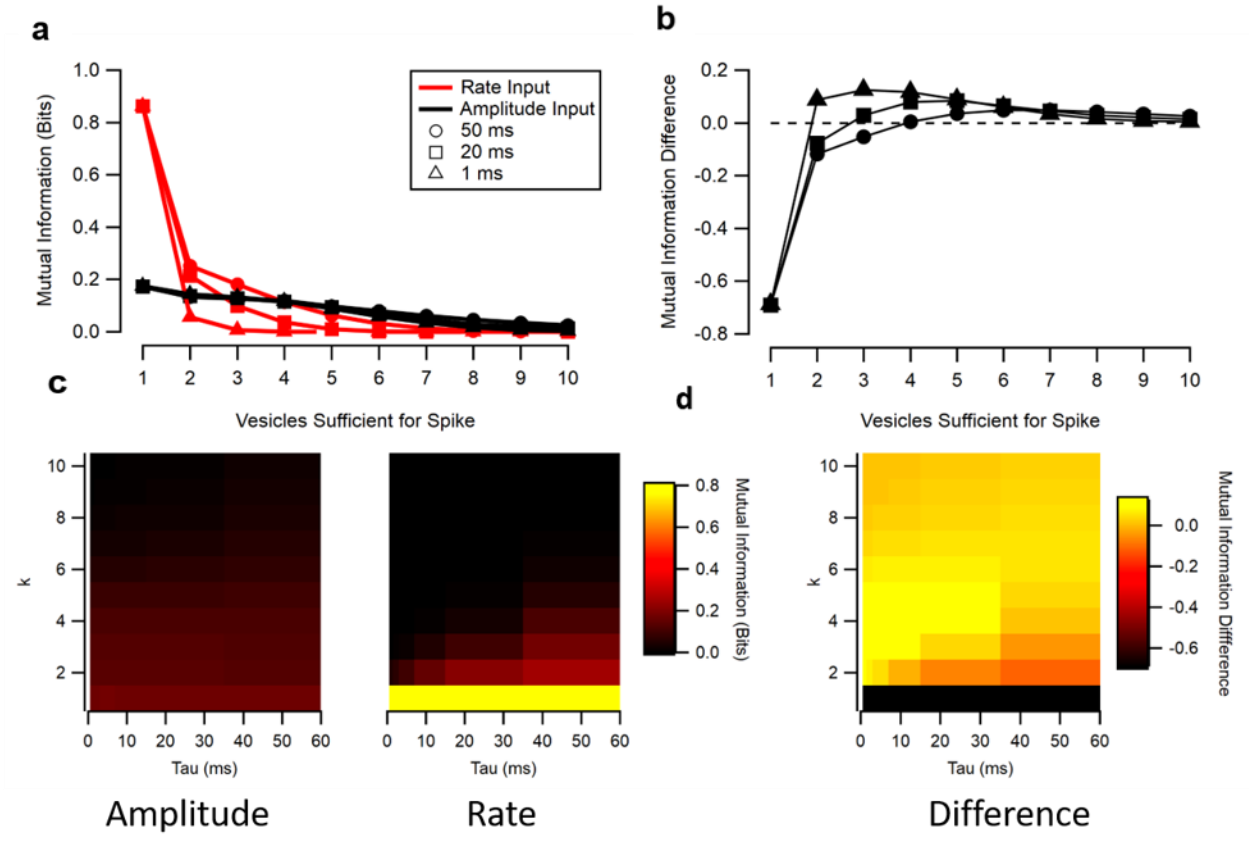

Figure 5.7: Spike Sequence Information for Amplitude (black) and Rate (red) inputs as a function of tau and k.

a) Mutual information for the rate input (dash) and amplitude input (solid) as a function of the vesicles sufficient to generate a spike. Values shown for tau of 50, 20, 10, and 1 ms. **b)** Mutual information difference (amplitude – rate) for the data shown in A. **c)** Mutual information (as a) in matrix form. **d)** Mutual information difference in matrix form.

It is notable that the stimulus distribution used in this work was only varied as a function of contrast – not frequency. It is thus not surprising that more information would be conveyed in the distribution of spike counts than spike times/sequences.

5.3.8: Increasing Inputs

Up to this point, we have considered neurons receiving monosynaptic input. While this situation does occur in certain areas of the nervous system (such as the synapse between the IHC and ANF), it is hardly the most common wiring. In general neurons receive input from many neurons, and the postsynaptic neuron integrates information both in time (from multiple events arising from a single input) as well as space (from multiple inputs).

In order to further study how this might affect information processing between either rate or amplitude coded inputs, we increased the number of inputs from one to ten, while simultaneously increasing the number of vesicles sufficient to generate a spike. Note that it was not necessary to do so – we could have simulated all pairs of k and inputs. However, this would exponentially increase the number of simulations, as well as produce many sets of situations with ‘less-than-believable’ physiological properties – i.e.. spike rates of 1 kHz without the addition of a refractory period. In equalizing the two variables, we do introduce another interesting property – ignoring leak and conductance properties of the cell, the total current input to the cell between all cases is identical. Thus, a simple perfect-integrator current-based model would produce an equal mean spike rate at any k/input conditions. This gives us the opportunity of comparing how either the rate case or amplitude case can transmit information while roughly equalizing the expected spike outputs for any value of k/inputs. We can thus see how the introduction of leak and conductance affect information rates.

In the previous sections, we observed how the spiking output of either the rate or the amplitude condition affects spike generation and information transmission when a single vesicle is able to generate a spike. This is identical to our current construction of k/in = 1. We then explored the specific case of when k = 5. Now, we can re-examine that question, but instead assuming that we have five times as many inputs (k = 5 = inputs). To aid in illustrating the differences, we show in **Fig5.4** the contrast response functions and spike count variance/mean plots for the three conditions mentioned above. We can begin by noting the differences between

our earlier k = 1, inputs = 1 condition and this one (**Fig 5.4**). Consistent with a total increase in arriving vesicles we see an increase in spike rate. Again, the rate condition is far more affected by the leak than the amplitude condition, although the amplitude condition is more affected in this case than in the k = 5, inputs = 1 case. Note that the 50 ms leak rate case is only slightly more effective at generating spikes than the amplitude case of the same leak at high contrasts; however, the rate condition has significantly lower variance – approaching the levels of an Erlang process. Thus in a perfect integrator, as the rate of vesicles input to the system increases concurrently with the number of vesicles that generate depolarization sufficient to generate a spike, we would expect the mean spike count to remain roughly equal across simulated values of k/inputs (roughly because some spikes will be lost due to leak), while the variability in the spike count to decrease as a consequence of the input statistics. This equates to decreasing conditional entropy while maintaining unconditional entropy, resulting in a higher value of mutual information. Notably, understanding both the slight decrease in interspike interval (as evidenced by more spikes) and decrease in variability makes it clearer why the Gamma distribution (the non-integer analogue of the Erlang distribution) is commonly seen in neuroscience. Input statistics could make the cell operate by an Erlang distribution, but the loss of voltage due to leak shifts the distribution to lower, and less variable, values.

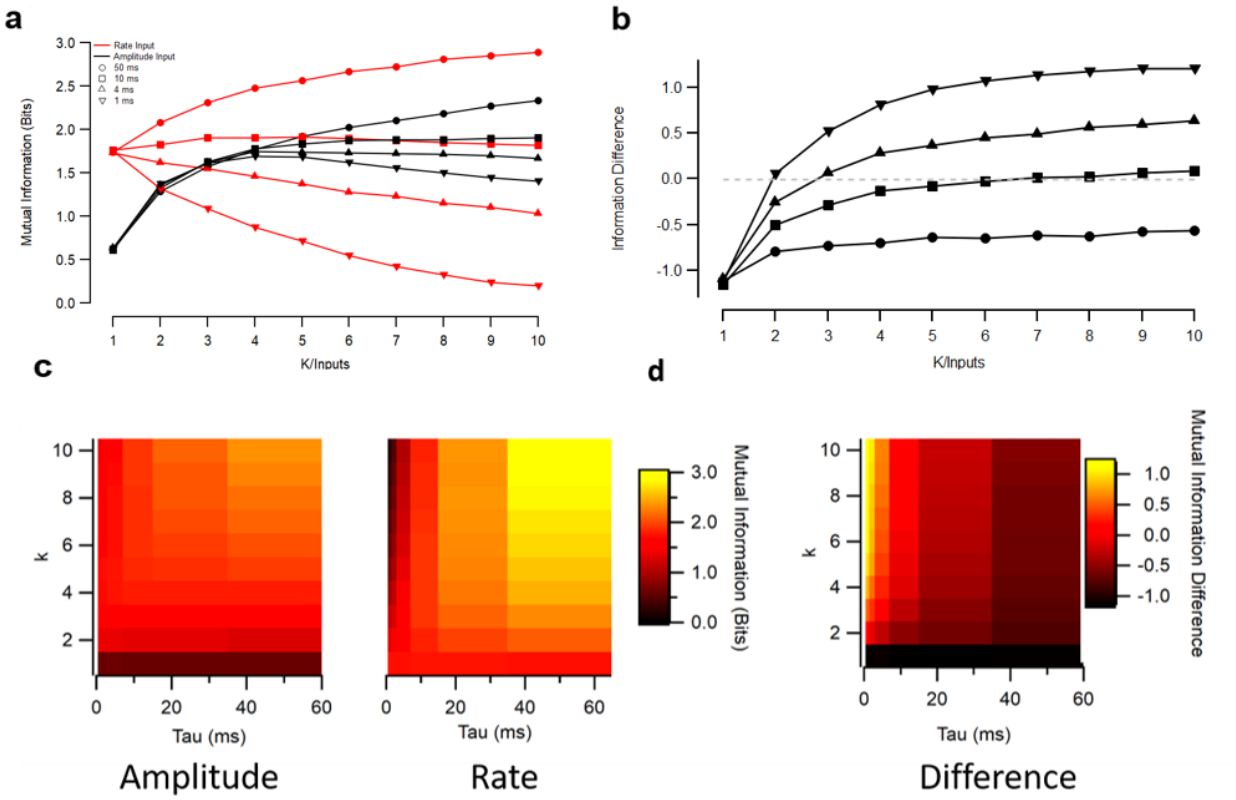

Figure 5.8: Spike count mutual information as a function of tau and k/inputs.

a) Mutual information for rate (red) and amplitude (black) as a function of tau and k/inputs. **b)** Mutual information difference (amplitude – rate) for the same data. **c)** Matrix plots of the data in a. **d)** Matrix form of the data in b.

To test this, we again computed the mutual information for each case, and plotted the results in **Fig5.10**. Perhaps surprisingly, we see that with the rate input condition, an increase in k (when k=inputs) no longer always decreases the mutual information (**Fig5.8a**) – when the membrane time constant τ is slow enough relative to the increase in input vesicles, the leak is overcome and spikes are generated, resulting in a net increase in mutual information. However, when the leak is fast enough, the increase in input vesicles cannot overcome it – and the spike rate and information content is decreased. On the other hand, the hybrid input condition (**Fig5.8a**) starts with increasing information as a function of k for all values of tau, and only begins to decrease later if the leak is sufficiently fast. Why might this be the case? Note that for the leak values at which this occurs, we find the maximum information content corresponding to a value of k approximately 4-5. As previously stated, based on our model formulation, this number of vesicles is one that can be easily reached in the amplitude code for higher contrasts

– an amplitude coded cell receiving high contrast input can be expected to produce events consisting of at least 5 vesicles – sufficient to generate a spike regardless of the leak value. As this value k increases, it becomes less likely that any individual input will produce a single event sufficient to cause the cell to spike. Thus, it would require multiple cells releasing high quantal events to generate a spike. Whether or not these multiple inputs will effectively integrate in the postsynaptic cell is then dependent upon the leak constant τ. Like the rate coded input, slower tau values allow for temporal integration of longer time scales – and the information content can continue to increase in the hybrid case. When the leak is fast, multiple inputs cannot be efficiently integrated (regardless of their quantal content), and spike generation and information capabilities decrease.

This brings us to an interesting point – whether or not increasing the inputs (while simultaneously increasing the vesicles sufficient to induce enough depolarization to evoke a spike) is able to increase information content depends both upon the coding strategy used as well as the cells leak constant. When the rate of incoming events is high relative to the leak constant tau, events from multiple inputs will be effectively summated in the postsynaptic cell. When the rate of incoming events is low relative to the leak constant, too much voltage will be lost due to leak between successive events and spikes are not efficiently generated. In **Fig5.10b**, we plotted the difference in information (Amplitude-rate) as a function of k/inputs. For slower leak values, the rate case dominates, while for faster leak values, the amplitude case is more advantageous. Here, we plotted values for k/inputs up to ten vesicles. From the graph, it would seem that as we increase these values that the difference in information transmission between rate and amplitude case seem to converge to a non-zero value, dependent upon the membrane time constant.

However, extending this analysis to k/inputs = 50 vesicles shows a different picture, where the information difference converges to zero for all values of tau. Thus, if one continues to increase the number of inputs to a cell, at a certain point the time-series of vesicles input to the system will become indistinguishable between rate or amplitude case. For example, if a million inputs are releasing vesicles at a constant rate of 100 vesicles per second (in either rate-modulated or amplitude-modulated fashion), it would be nearly impossible for a postsynaptic cell to distinguish between a single amplitude-coded event consisting of six vesicles, or six of the one million inputs released vesicles ‘nearly simultaneously’. Because the ‘temporal resolution’ of the postsynaptic set based upon the leak constant tau, with so many inputs the postsynaptic cell ‘wouldn’t care’ whether the inputs were amplitude or rate modulated. Thus, the differences in

spike output between the rate and hybrid conditions are minimized, and the mutual information differences converges to zero.

5.4: Discussion

Sensory cells and circuits have been adapted for highly complex and specific computations, often thought to achieve specific goals or optimize specific problems (Barlow 1961, Olshausen and Field 1996, Schwartz and Simoncelli 2001, Harper and McAlpine 2004). Here, we constructed a statistical framework with which we can compare how information transmission in neural systems might be affected by a rate or hybrid-based code. The approach to modeling MVR as a Poisson Process is not new – previous work has attempted to model the phenomenon as a Compound Poisson Process (CPP) (Neef et al, 2007). However, this approach does not allow for time-dependence in MVR. To allow for this, we have introduced the use of time-dependent Poisson Splitting. Thus, MVR may be more accurately modelled, and its potential effects on postsynaptic cells may be illuminated. Utilizing this process, we can equalize the number of vesicles released at any given time between a rate code, or an alternative amplitude or hybrid-based code. In doing so, we have described in which physiological conditions MVR may be beneficial to the nervous system, and how this switch from a rate code to an alternative code may impact information processing.

We show that a switch from a rate-based code to a hybrid code can allow for increased information transmission, and this effect is highly dependent upon physiological properties of the post-synaptic cell. Cells with a middling input resistance and fewer inputs are able to transmit more information using a hybrid code than an amplitude code, while cells with very high resistance, such that single vesicles can generate a spike, transmit more information using a rate-based code. From this, we might then expect that MVR might be more commonly found in cells with these sorts of physiological parameters. Thus, midget RGCs or those found near the fovea, which have fewer inputs than RGCs found at higher eccentricities, might be expected to utilize MVR more often – or at least more effectively. However, this is clearly not always applicable. MVR in IHCs, for example, is a ubiquitous property, despite the fact that cells postsynaptic have resistances allowing for a single vesicle to generate a spike. Here, we have shown that MVR might be detrimental to information transmission in cells like this. This then raises the question of why would the nervous system then use MVR at this synapse? In this work, we focused on processing of temporal contrast in order to mimic experimental data

recorded in BC glutamate release. While we have attempted to make this analysis as system-agnostic as possible, it might simply be that the types of information being processed by IHCs in the early auditory system might differ from those of the retina. Our choice of stimulus could then affect the analysis – perhaps using a stimulus in which the frequency of the stimulus is altered, rather than the contrast, would result in differences in information processing.

Regardless of choice of stimuli, we have developed a manner in which to simulate time-dependence in MVR. We can then create a corresponding rate code with which we can compare how information can be transmitted. Here, we hope to realize a new framework for understanding the implications of MVR and hybrid coding in the nervous system. In the future, we hope this framework will allow for additional understanding of synaptic transmission in the nervous system.

Chapter 6: Conclusions

Within this work, I aimed to demonstrate the possible utility of multivesicular release in the sensory systems. In doing so, the overarching goal was to present a mechanism for information transmission in neural systems in order to further illuminate the neural code. Information in neural systems can be represented in three distinct – yet overlapping – regimes. The traditional binary spike, continuous modulation of membrane potential, and now a 'multinary' signal. In doing so, multiple new areas of research involving amplitude coding and MVR become available.

6.1: A New Technique for Quantizing Vesicle Release

The first aim of this work was to develop a manner in which *in vivo* optical recordings of glutamate release can be quantized into units of individual vesicles. While the initial discoveries of multivesicular release came about decades ago (Tong and Jahr 1994, Auger, Kondo et al. 1998, von Gersdorff, Sakaba et al. 1998), the bulk of the work on the phenomenon utilized electrophysiological or electrochemical recordings – techniques which necessitate direct contact with the recorded cell and increasing the likelihood of either damaging circuitry connections in the organism or picking up signals from multiple cells, making a functional study of MVR difficult. The techniques presented in this work – Wiener deconvolution and Gaussian Mixture Models (GMM) – have existed for decades (McLachlan 2000). Notably, while Wiener deconvolution has been used in the analysis of MVR, I am the first in my knowledge to combine this with a GMM. Adding this second step allows for a statistical analysis of quantization of vesicular release, realizing the ability to count vesicles.

6.2: Information Capacity of MVR

The second main goal of this work was to demonstrate that MVR can be used in an informative way. As the initial findings of MVR had focused little on the functional properties of MVR, here I focused on the potential functional benefits. By stimulating BCs with a variety of temporal contrasts – one of the most basic elements of the visual scene – I found that MVR events are preferentially evoked by higher contrasts, as shown in both the Transmitter Triggered Average as well as the shift of distribution in quanta per event at higher contrasts. This tendency was then quantified using information theoretical values. As the specific information (I2) per vesicle increased with quantal content I conclude that higher quantal events have the capacity to transmit information more efficiently than lower quantal events. Information theoretical metrics are notoriously difficult to estimate, however. While precautions were taken to ensure no bias

was included in our analysis, the truth remains that these are *estimates*. Due to inherent physical properties of imaging, it is difficult to produce recordings of any single BC that match the lengths of electrophysiological recordings. Thus, it is important to note the raw values of the information theoretical estimates will likely change as technology improves and longer recordings can be taken. However, the focus of this analysis was not on the values themselves, but on the trend of higher quantal events to contain more information per vesicle than lower quantal events. As this was found in the majority of cells analyzed, I am confident of the veracity of the results – while the exact information values may change with further research, the trend will likely remain.

6.3. Comparisons to Other Work

Analysis concerning the potential mechanisms for information transmission between neurons are far from rare (Theunissen and Miller 1991, de Ruyter van Steveninck and S.B. 1996, Juusola, French et al. 1996, Borst and Theunissen 1999, Field and Chichilnisky 2007, Foffani, Morales-Botello et al. 2009). While this work was experimentally focused on the visual system of vertebrates, the visual system of insects is also highly studied, with numerous similarities between the two systems. In the insect retina, light is first detected by sets of six photoreceptors (PRs) which synaptically transmit the visual signal to large monopolar cells (LMCs) of the lamina (Laughlin 1987, Laughlin 1989, Sanes and Zipursky 2010). Like in the vertebrate retina, the PRs of the insect retina respond to the visual environment with the isomerization of rhodopsin molecules, resulting in a ‘bump’ of depolarization within the cell. Most striking about the system is its ability to adapt to ambient light levels – quantal bumps occurring in darkness result in a membrane depolarization of one or more mV, while under daylight conditions quantal bumps fuse together in time and reduce their amplitude, resulting in membrane depolarization ins the μV ranges (Laughlin 1987, Laughlin, Howard et al. 1987, Zheng, Nikolaev et al. 2009). Like in the quantal bumps of insect PRs, zebrafish BCs appear to operate in a discrete manner - releasing quantal amounts of neurotransmitter according to a stochastic process (Yeandle and Spiegler 1973), with BCs releasing glutamate and insect PRs releasing histamine (Katz and Minke 2009, Sanes and Zipursky 2010). However, while at higher illumination levels the insect quantal bumps fuse and become continuous, the quantal release of vesicles from BC terminals remain visibly discrete, allowing for quantal decomposition even at high release rates (Laughlin 1989, Singer, Lassova et al. 2004, James, Darnet et al. 2019).

While not examined at the level of the synapse, the use of multiple symbols in neural information transmission is not novel. In particular, it has been shown that thalamocortical

neurons are capable of spike bursts – rapid successions of action potentials (Zeldenrust, Chameau et al. 2013, Zeldenrust, Chameau et al. 2018). By increasing the temporal window of analysis, the number of spikes occurring in a burst can be seen as a non-binary, discrete metric (Longden, Wicklein et al. 2017, Zeldenrust, Chameau et al. 2018, Zeldenrust, Wadman et al. 2018), analogous to how MVR has been presented here. The work presented here differs in two main regards. Spike analyses are conducted at the level of the action potential, not at the level of the synapse as described here, and thus the stochastic nature of vesicle release is not reflected in the analysis. Additionally, in spike burst analysis, the time windows used in analysis are much longer than those used here. In fact, it is likely that further technological advancements will allow for a further decrease in the time scales for theoretical analysis of MVR, as MVR can be coordinated down to μs timescales (Singer, Lassova et al. 2004)

6.4. Effects of Changing Base

How does switching from a binary system to a multinary system affect computation and information processing? Consider a binary system (such as the binary activity of a set number of neurons, or the state of a binary switch) encoding an integer using n values. First, note that in order to represent an integer in a given base, one uses the equation $N = (b-1)(b^0) + (b-1)(b^1) + \cdots + (b-1)(b^n - 1)$, where N is the integer being encoded, b is the base used, and n is the total number of elements. Thus, in any system, one can calculate the maximum integer the system can encode by finding the sum of this series: $N = (b-1)\left(\frac{1-b^n}{1-b}\right) = b^n - 1$. For a standard 8-bit system (where integers are encoded by the binary status of 8 elements), the maximum integer that can be encoded is $2^n - 1 = 255$. Now, consider a different system with three possible states – this could reference the simultaneous release of zero, one, or two vesicles. Within this system, a set of 8 elements can encode $N = 3^8 - 1 = 6560$, an increase of over 25-fold. Thus, in either a computer-scientific or neuroscientific basis, more information can be encoded by increasing the number of possible states from two to three.

Perhaps a better way to phrase this in the neuroscientific terms is: how much time (bins of neural code) can we save by switching from one base to another? Encoding ten bits of information using a binary system would take ten bins, while a system that goes to *eleven* symbols could achieve this in a single bin, ten percent of the amount of time required for a binary system.

Why, then, don't all computational devices (synthetic or neural) not use a ternary – or even higher – base system? The answer, quite simply, is noise. While studies have shown that using a base 3 computer system can more efficiently operate, requiring less time and energy than a base 2 system, a ternary system is far more susceptible to noise than a binary system. For a simple two state system, the default states are off and on (corresponding to no activity or activity). Thus, binary systems can set a threshold to what is considered activity – this prevents any small activity due to noise being incorrectly labeled as activity. The ternary system, on the other hand, must indicate more than just an on/off signal – it requires either a difference in magnitudes between the values corresponding to one and two in the most basic ternary system, or the ability to switch sign (using the values of -1, 0 , and 1) for what is referred to as a balanced ternary system. Thus, while noise is unlikely to affect the status of a binary system, a ternary system requires more sophisticated hardware capable of differentiating between multiple signals. Consequently, noise values that might not interfere with the operation of a binary computer system may induce errors in a ternary system. Put in neuroscientific terms – if a neuron is using MVR to encode values from 0 to n, the postsynaptic cell receiving this neuron's synaptic input must be capable of reliably differentiating the depolarization induced from each possible numbers of vesicles released. Cells receiving input from traditional spiking cells, on the other hand, require only the machinery necessary to detect the differences between no input activity and input activity. Neurons that take input from multinary systems, then, are more susceptible to noise than neurons taking input from binary systems.

6.5: More Realistic Models

The model of MVR presented in this work is not the most biologically accurate – it was designed not in order to model MVR with maximal accuracy, but rather to allow for a manner in which the effects of MVR can be statistically compared to traditional models of vesicle release. As such, aspects such as adaptation or vesicle release dynamics are lacking in the presented model. However, constructing more realistic models allowing for MVR is possible, and we can proceed in one of two perspectives: the biologically driven, or the theoretically driven.

How can we implement MVR in more realistic theoretical models? Specifically, we would like to model the process from the true input – visual stimuli. In fact, many of the techniques which we can used to do so already exist and simply require a bit of reworking. To begin, consider the Linear-Nonlinear Bernoulli model (Williamson, Sahani et al. 2015). Here, a cell is

characterized by its receptive field, a linear kernel, $\vec{k}$. The primary excitatory drive to the models arises from the projection of the stimulus $\vec{x}(t)$ onto the cells receptive field. The output of this linear convolution is then passed through a nonlinearity, resulting in the probability of a Bernoulli event:

$$p = n(\vec{k} \cdot \vec{x}(t))$$

While the traditional LNB model stops here – the result is either a spike, or the absence of a spike – we can add an additional compounding function that incorporates a history effect mimicking the dynamics of vesicle release as well as the ability to alter the amplitude of an event based on recent stimuli:

$$q|ev \sim g(\vec{k} \cdot \vec{x}(t) - \sum_{j=0}^{i-1} h(t, t_j, q_j)$$

Here, the probability of a quanta given the presence of an event, q, is distributed according to some linear projection of the stimulus onto the receptive field kernel, with a history filter that incorporates vesicle release in the recent history.

While initially this model might not seem much different from the traditional LNP model – both can allow for varying amplitudes and history effects – take note that these models diverge when the bin size decreases. While the LNP model reduces to an LNB model altering the rate of single events, this model continues to show MVR behavior even as the bin size approaches zero. The timescales with which MVR operates in early sensory systems, notably, are considerably smaller than those offered by traditional models, and as such LNP and GLM models do not perform well in modelling MVR. This model could allow for the refinement necessary to accurately model MVR as well as the statistical ease of traditional neural models – traditional methods of gradient ascent should work in this case, although they will be complicated by the multidimensional history filter necessary to capture dynamics of vesicle release.

The alternative perspective we can view models of MVR is the more biologically driven. As these models – as opposed to the earlier theoretical models - often contain remarkably complex dependencies, they are often capable of describing arbitrarily complex biological processes, at the cost of abolishing any traditional method of statistical parameter inference. Nonetheless, work I completed with colleagues allowed for a mechanistic model of vesicle release to be fit use approximate Bayesian inference. However, further development of models in this vein is difficult, as the amount of data that can be reliably collected due to physical aspects of imaging is highly limited compared to electrophysiological methods (Schroeder, James et al. 2019).

6.6: MVR and Multiplexing, Adaptation

Amplitude coding and MVR could potentially drastically improve the nervous systems information processing ability by encoding the adaptive state of the neuron separately from the stimulus state. In the present work adaptation was not only not discussed, but also actively avoided – adaptive effects complicate information theoretic measurements by inducing history effects, so short stimuli were predominately used to avoid them. However, initial analyses indicate that in some cells adaptation may preferentially affect the distribution of event amplitude versus event rates. As an information processing system, this ‘adaptive multiplexing’ can allow for the ability of a postsynaptic cell to disentangle a presynaptic cells adaptive state from its stimulus drive without explicit history knowledge itself. Rather than the postsynaptic cell having to ‘remember’ its input history in order to disentangle adaptation versus contrast, it can simply use the rate of events as a readout of stimulus contrast and the event amplitude as a metric of the presynaptic cells input history.

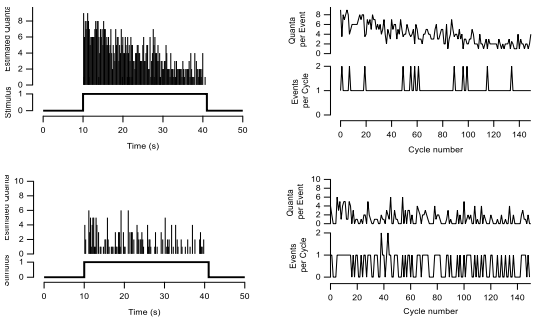

Figure 6.1: An adaptive multiplexing cell

a) Estimated quantal time series for a cell in response to a 100% full field 5 Hz stimulus. **b)** The average quanta per event for each of the 150 cycles of the stimulus in the **a)** and the corresponding number of events in each cycle. Note that this synapse reliably releases single events on all but a few trials at this contrast, regardless of adaptive state, but the average number of vesicles is reduced in in time. **c)** The same cell in **a),** but presented with 60% contrast. **d)** Corresponding quanta per event and event per cycle plot for the data in **c)**. Here, the lower contrasts triggers fewer events as well as events with fewer vesicles. Note that despite the adaptation present in **b)**, the desensitized response to 100% contrast is distinguishable from the response to 60% contrast, as this cell adapts by reducing quantal content and not event frequency.

Notably, this can potentially allow for increasing SNR even after prolonged high contrasts. In traditional rate driven models, the event rate after prolonged stimulation reduces and, ignoring disinhibition, reductions in contrast are poorly encoded by further reducing event rate from an already low value. In the multiplexed adaptive cell, where adaptation reduces the event amplitude, but not rate, reductions in contrast can be more simply signaled, as the event rate remains high even after adaptation – only the event amplitude decreases as a function of adaptive state. See **Fig6.1** for an example from an experimental recording.

References

Abbott, L. F. (1999). "Lapicque's introduction of the integrate-and-fire model neuron (1907)." Brain Res Bull **50**(5-6): 303-304.

Abbott, L. F. and W. G. Regehr (2004). "Synaptic computation." Nature **431**(7010): 796-803.

Ahrens, M. B. (2019). "Zebrafish Neuroscience: Using Artificial Neural Networks to Help Understand Brains." Curr Biol **29**(21): R1138-R1140.

Antinucci, P., O. Suleyman, C. Monfries and R. Hindges (2016). "Neural Mechanisms Generating Orientation Selectivity in the Retina." Curr Biol **26**(14): 1802-1815.

Atick, J. J. and A. N. Redlich (1992). "What does the retina know about natural scenes?" Neural computation **4**(2): 196-210.

Auger, C., S. Kondo and A. Marty (1998). "Multivesicular release at single functional synaptic sites in cerebellar stellate and basket cells." J Neurosci **18**(12): 4532-4547.

Baden, T., T. Euler and P. Berens (2020). "Understanding the retinal basis of vision across species." Nat Rev Neurosci **21**(1): 5-20.

Baden, T., T. Euler, M. Weckstrom and L. Lagnado (2013). "Spikes and ribbon synapses in early vision." Trends Neurosci **36**(8): 480-488.

Baden, T., A. Nikolaev, F. Esposti, E. Dreosti, B. Odermatt and L. Lagnado (2014). "A synaptic mechanism for temporal filtering of visual signals." PLoS Biol **12**(10): e1001972.

Baden, T. and D. Osorio (2019). "The Retinal Basis of Vertebrate Color Vision." Annu Rev Vis Sci **5**: 177-200.

Balasubramanian, V., D. Kimber and M. J. Berry, 2nd (2001). "Metabolically efficient information processing." Neural Comput **13**(4): 799-815.

Barlow, H. B. (1961). "Possible principles underlying the transformation of sensory messages." Sensory communication **1**: 217-234.

Berry, M. J., D. K. Warland and M. Meister (1997). "The structure and precision of retinal spike trains." Proc Natl Acad Sci U S A **94**(10): 5411-5416.

Borst, A. and F. E. Theunissen (1999). "Information theory and neural coding." Nat Neurosci **2**(11): 947-957.

Brand, A. H. and N. Perrimon (1993). "Targeted gene expression as a means of altering cell fates and generating dominant phenotypes." development **118**(2): 401-415.

Burkitt, A. N. (2006). "A review of the integrate-and-fire neuron model: I. Homogeneous synaptic input." Biol Cybern **95**(1): 1-19.

Burkitt, A. N. (2006). "A review of the integrate-and-fire neuron model: II. Inhomogeneous synaptic input and network properties." Biol Cybern **95**(2): 97-112.

Burrone, J. and L. Lagnado (2000). "Synaptic depression and the kinetics of exocytosis in retinal bipolar cells." J Neurosci **20**(2): 568-578.

Calhoun, A. J., J. W. Pillow and M. Murthy (2019). "Unsupervised identification of the internal states that shape natural behavior." Nat Neurosci **22**(12): 2040-2049.

Carrillo-Medina, J. L. and R. Latorre (2018). "Detection of Activation Sequences in Spiking-Bursting Neurons by means of the Recognition of Intraburst Neural Signatures." Sci Rep **8**(1): 16726.

Chamberland, S., A. Evstratova and K. Toth (2014). "Interplay between synchronization of multivesicular release and recruitment of additional release sites support short-term facilitation at hippocampal mossy fiber to CA3 pyramidal cells synapses." J Neurosci **34**(33): 11032-11047.

Charles, A. S., M. Park, J. P. Weller, G. D. Horwitz and J. W. Pillow (2018). "Dethroning the Fano Factor: A Flexible, Model-Based Approach to Partitioning Neural Variability." Neural Comput **30**(4): 1012-1045.

Chatfield, C. and G. J. Goodhardt (1973). "A Consumer Purchasing Model with Erlang Inter-Purchase Time." Journal of the American Statistical Association **68**.

Chiu, S., D. Stoyan, W. Kendall and J. Mecke (2013). Stochastic Geometry and Its Applications.
Cho, S. and H. von Gersdorff (2012). "Ca(2+) influx and neurotransmitter release at ribbon synapses." Cell Calcium **52**(3-4): 208-216.
Christie, J. M. and C. E. Jahr (2006). "Multivesicular release at Schaffer collateral-CA1 hippocampal synapses." J Neurosci **26**(1): 210-216.
Cinlar, E. (2013). Introduction to Stochastic Processes, Dover Publications, Incorporated.
Dayan, P. and L. F. Abbott (2001). Theoretical neuroscience: computational and mathematical modeling of neural systems, Computational Neuroscience Series.
de Ruyter van Steveninck, R., W. Bialek and H. B. Barlow (1988). "Real-time performance of a movement-sensitive neuron in the blowfly visual system: coding and information transfer in short spike sequences." Proceedings of the Royal Society of London. Series B. Biological Sciences **234**(1277): 379-414.
de Ruyter van Steveninck, R. R. and L. S.B. (1996). "The rate of information transfer at a graded-potential synapse." Nature **379**.
Del Castillo, J. and B. Katz (1954). "Quantal components of the end-plate potential." J Physiol **124**(3): 560-573.
Demb, J. B., K. Zaghloul, L. Haarsma and P. Sterling (2001). "Bipolar cells contribute to nonlinear spatial summation in the brisk-transient (Y) ganglion cell in mammalian retina." J Neurosci **21**(19): 7447-7454.
Dempster, A., N. Laird and D. Rubin (1977). "Maximum Likelihood from Incomplete Data via the EM Algorithm." Journal of the Royal Statistical Society.
DeWeese, M. R. and M. Meister (1999). "How to measure the information gained from one symbol." Network **10**(4): 325-340.
Dobrunz, L. E. and C. F. Stevens (1997). "Heterogeneity of release probability, facilitation, and depletion at central synapses." Neuron **18**(6): 995-1008.
Easter, S. S., Jr. and G. N. Nicola (1996). "The development of vision in the zebrafish (Danio rerio)." Dev Biol **180**(2): 646-663.
Eckhorn, R. and B. Popel (1975). "Rigorous and extended application of information theory to the afferent visual system of the cat. II. Experimental results." Biol Cybern **17**(1): 71-77.
Field, G. D. and E. J. Chichilnisky (2007). "Information processing in the primate retina: circuitry and coding." Annu Rev Neurosci **30**: 1-30.
Foffani, G., M. L. Morales-Botello and J. Aguilar (2009). "Spike timing, spike count, and temporal information for the discrimination of tactile stimuli in the rat ventrobasal complex." J Neurosci **29**(18): 5964-5973.
Foster, K. A., J. J. Crowley and W. G. Regehr (2005). "The influence of multivesicular release and postsynaptic receptor saturation on transmission at granule cell to Purkinje cell synapses." J Neurosci **25**(50): 11655-11665.
Franke, K., P. Berens, T. Schubert, M. Bethge, T. Euler and T. Baden (2017). "Inhibition decorrelates visual feature representations in the inner retina." Nature **542**(7642): 439-444.
Freed, M. A. (2005). "Quantal encoding of information in a retinal ganglion cell." J Neurophysiol **94**(2): 1048-1056.
Fuchs, P. A. (2005). "Time and intensity coding at the hair cell's ribbon synapse." J Physiol **566**(Pt 1): 7-12.
Fuchs, P. A. and E. Glowatzki (2015). "Synaptic studies inform the functional diversity of cochlear afferents." Hear Res **330**(Pt A): 18-25.
Gautrais, J. and S. Thorpe (1998). "Rate coding versus temporal order coding: a theoretical approach." Biosystems **48**(1-3): 57-65.
Gelman, A., J. B. Carlin, H. S. Stern, D. B. Dunson, A. Vehtari and D. B. Rubin (2013). Bayesian Data Analysis, Third Edition, Taylor & Francis.

Gerstein, G. L. and W. A. Clark (1964). "Simultaneous Studies of Firing Patterns in Several Neurons." Science **143**(3612): 1325-1327.
Gerstner, W., A. K. Kreiter, H. Markram and A. V. Herz (1997). "Neural codes: firing rates and beyond." Proc Natl Acad Sci U S A **94**(24): 12740-12741.
Gibbs, J. W. (2015). Elementary Principles in Statistical Mechanics, Developed with Especial Reference to the Rational Fo - Scholar's Choice Edition, Creative Media Partners, LLC.
Gilles, J. F., M. Dos Santos, T. Boudier, S. Bolte and N. Heck (2017). "DiAna, an ImageJ tool for object-based 3D co-localization and distance analysis." Methods **115**: 55-64.
Glowatzki, E. and P. A. Fuchs (2002). "Transmitter release at the hair cell ribbon synapse." Nat Neurosci **5**(2): 147-154.
Gollisch, T. and M. Meister (2008). "Rapid neural coding in the retina with relative spike latencies." Science **319**(5866): 1108-1111.
Gollisch, T. and M. Meister (2010). "Eye smarter than scientists believed: neural computations in circuits of the retina." Neuron **65**(2): 150-164.
Gomis, A., J. Burrone and L. Lagnado (1999). "Two actions of calcium regulate the supply of releasable vesicles at the ribbon synapse of retinal bipolar cells." The Journal of neuroscience : the official journal of the Society for Neuroscience **19**(15): 6309-6317.
Goris, R. L., J. A. Movshon and E. P. Simoncelli (2014). "Partitioning neuronal variability." Nat Neurosci **17**(6): 858-865.
Goutman, J. D. and E. Glowatzki (2011). "Short-term facilitation modulates size and timing of the synaptic response at the inner hair cell ribbon synapse." J Neurosci **31**(22): 7974-7981.
Grabner, C. P. and T. Moser (2018). "Individual synaptic vesicles mediate stimulated exocytosis from cochlear inner hair cells." Proc Natl Acad Sci U S A **115**(50): 12811-12816.
Grabner, C. P. and D. Zenisek (2013). "Amperometric resolution of a prespike stammer and evoked phases of fast release from retinal bipolar cells." J Neurosci **33**(19): 8144-8158.
Gundelfinger, E. D., C. Reissner and C. C. Garner (2015). "Role of Bassoon and Piccolo in Assembly and Molecular Organization of the Active Zone." Front Synaptic Neurosci **7**: 19.
Halpern, M. E., J. Rhee, M. G. Goll, C. M. Akitake, M. Parsons and S. D. Leach (2008). "Gal4/UAS transgenic tools and their application to zebrafish." Zebrafish **5**(2): 97-110.
Han, W., L. A. Tellez, M. J. Rangel, Jr., S. C. Motta, X. Zhang, I. O. Perez, N. S. Canteras, S. J. Shammah-Lagnado, A. N. van den Pol and I. E. de Araujo (2017). "Integrated Control of Predatory Hunting by the Central Nucleus of the Amygdala." Cell **168**(1-2): 311-324 e318.
Harper, N. S. and D. McAlpine (2004). "Optimal neural population coding of an auditory spatial cue." Nature **430**(7000): 682-686.
Heuser, J. E. and T. S. Reese (1973). "Evidence for recycling of synaptic vesicle membrane during transmitter release at the frog neuromuscular junction." J Cell Biol **57**(2): 315-344.
Higley, M. J., G. J. Soler-Llavina and B. L. Sabatini (2009). "Cholinergic modulation of multivesicular release regulates striatal synaptic potency and integration." Nat Neurosci **12**(9): 1121-1128.
Hjelmstad, G. O., R. A. Nicoll and R. C. Malenka (1997). "Synaptic refractory period provides a measure of probability of release in the hippocampus." Neuron **19**(6): 1309-1318.
Hodgkin, A. L. and A. F. Huxley (1952). "A quantitative description of membrane current and its application to conduction and excitation in nerve." J Physiol **117**(4): 500-544.
Huang, C. H., J. Bao and T. Sakaba (2010). "Multivesicular release differentiates the reliability of synaptic transmission between the visual cortex and the somatosensory cortex." J Neurosci **30**(36): 11994-12004.
Huxter, J., N. Burgess and J. O'Keefe (2003). "Independent rate and temporal coding in hippocampal pyramidal cells." Nature **425**(6960): 828-832.
James, B., L. Darnet, J. Moya-Diaz, S. H. Seibel and L. Lagnado (2019). "An amplitude code transmits information at a visual synapse." Nat Neurosci **22**(7): 1140-1147.

Johnson, D. H. (1996). "Point process models of single-neuron discharges." J Comput Neurosci **3**(4): 275-299.
Johnston, J., S. H. Seibel, L. S. A. Darnet, S. Renninger, M. Orger and L. Lagnado (2019). "A Retinal Circuit Generating a Dynamic Predictive Code for Oriented Features." Neuron **102**(6): 1211-1222 e1213.
Juusola, M., A. S. French, R. O. Uusitalo and M. Weckstrom (1996). "Information processing by graded-potential transmission through tonically active synapses." Trends Neurosci **19**(7): 292-297.
Kandel, E. R., J. H. Schwartz and T. M. Jessell (2000). Principles of Neural Science, McGraw-Hill.
Karczmar, A. G. (1996). "The Otto Loewi Lecture. Loewi's discovery and the XXI century." Prog Brain Res **109**: 1-27, xvii.
Katz, B. and B. Minke (2009). "Drosophila photoreceptors and signaling mechanisms." Front Cell Neurosci **3**: 2.
Kay, L. M. and S. M. Sherman (2007). "An argument for an olfactory thalamus." Trends Neurosci **30**(2): 47-53.
Kingman, J. F. C. (1992). Poisson Processes, Clarendon Press.
Korn, H., A. Mallet, A. Triller and D. S. Faber (1982). "Transmission at a central inhibitory synapse. II. Quantal description of release, with a physical correlate for binomial n." J Neurophysiol **48**(3): 679-707.
Korn, H., C. Sur, S. Charpier, P. Legendre and D. S. Faber (1994). "The one-vesicle hypothesis and multivesicular release." Adv Second Messenger Phosphoprotein Res **29**: 301-322.
Lagnado, L. and F. Schmitz (2015). "Ribbon Synapses and Visual Processing in the Retina." Annu Rev Vis Sci **1**: 235-262.
Laughlin, S. B. (1987). "Form and function in retinal processing." Trends in Neurosciences **10**(11): 478-483.
Laughlin, S. B. (1989). "The role of sensory adaptation in the retina." J Exp Biol **146**: 39-62.
Laughlin, S. B. (2001). "Energy as a constraint on the coding and processing of sensory information." Curr Opin Neurobiol **11**(4): 475-480.
Laughlin, S. B., J. Howard and B. Blakeslee (1987). "Synaptic Limitations to Contrast Coding in the Retina of the Blowfly Calliphora." Proceedings of the Royal Society of London. Series B, Biological Sciences **231**(1265): 437-467.
Laughlin, S. B., J. Howard and B. Blakeslee (1987). "Synaptic limitations to contrast coding in the retina of the blowfly Calliphora." Proc R Soc Lond B Biol Sci **231**(1265): 437-467.
Lettvin, J. Y., H. R. Maturana, W. S. McCulloch and W. H. Pitts (1959). "What the Frog's Eye Tells the Frog's Brain." Proceedings of the IRE **47**(11): 1940-1951.
Levy, W. B. and R. A. Baxter (2002). "Energy-efficient neuronal computation via quantal synaptic failures." J Neurosci **22**(11): 4746-4755.
Li, G. L., S. Cho and H. von Gersdorff (2014). "Phase-locking precision is enhanced by multiquantal release at an auditory hair cell ribbon synapse." Neuron **83**(6): 1404-1417.
Lin, M. Z. and M. J. Schnitzer (2016). "Genetically encoded indicators of neuronal activity." Nat Neurosci **19**(9): 1142-1153.
Lisman, J. E., S. Raghavachari and R. W. Tsien (2007). "The sequence of events that underlie quantal transmission at central glutamatergic synapses." Nat Rev Neurosci **8**(8): 597-609.
London, M., A. Roth, L. Beeren, M. Hausser and P. E. Latham (2010). "Sensitivity to perturbations in vivo implies high noise and suggests rate coding in cortex." Nature **466**(7302): 123-127.
Longden, K. D., M. Wicklein, B. J. Hardcastle, S. J. Huston and H. G. Krapp (2017). "Spike Burst Coding of Translatory Optic Flow and Depth from Motion in the Fly Visual System." Curr Biol **27**(21): 3225-3236 e3223.
MacKay, D. M. and W. S. McCulloch (1952). "The limiting information capacity of a neuronal link." Bulletin of Mathematical Biophysics **14**: 127-135.

Maimon, G. and J. A. Assad (2009). "Beyond Poisson: increased spike-time regularity across primate parietal cortex." Neuron **62**(3): 426-440.
Marvin, J. S., B. G. Borghuis, L. Tian, J. Cichon, M. T. Harnett, J. Akerboom, A. Gordus, S. L. Renninger, T. W. Chen, C. I. Bargmann, M. B. Orger, E. R. Schreiter, J. B. Demb, W. B. Gan, S. A. Hires and L. L. Looger (2013). "An optimized fluorescent probe for visualizing glutamate neurotransmission." Nat Methods **10**(2): 162-170.
Masland, R. H. (2001). "The fundamental plan of the retina." Nat Neurosci **4**(9): 877-886.
McLachlan, G. J. (2000). Finite mixture models / Geoffrey McLachlan, David Peel. New York ; Chichester, Wiley.
Mennerick, S. and G. Matthews (1996). "Ultrafast exocytosis elicited by calcium current in synaptic terminals of retinal bipolar neurons." Neuron **17**(6): 1241-1249.
Moore, G. P., D. H. Perkel and J. P. Segundo (1966). "Statistical analysis and functional interpretation of neuronal spike data." Annu Rev Physiol **28**: 493-522.
Neef, A., D. Khimich, P. Pirih, D. Riedel, F. Wolf and T. Moser (2007). "Probing the mechanism of exocytosis at the hair cell ribbon synapse." J Neurosci **27**(47): 12933-12944.
Neves, G. and L. Lagnado (1999). "The kinetics of exocytosis and endocytosis in the synaptic terminal of goldfish retinal bipolar cells." J Physiol **515 (Pt 1)**: 181-202.
Nikolaev, A., K. M. Leung, B. Odermatt and L. Lagnado (2013). "Synaptic mechanisms of adaptation and sensitization in the retina." Nat Neurosci **16**(7): 934-941.
O'Brien, B. J., T. Isayama, R. Richardson and D. M. Berson (2002). "Intrinsic physiological properties of cat retinal ganglion cells." J Physiol **538**(Pt 3): 787-802.
O'Keefe, J. (1976). "Place units in the hippocampus of the freely moving rat." Exp Neurol **51**(1): 78-109.
Odermatt, B., A. Nikolaev and L. Lagnado (2012). "Encoding of luminance and contrast by linear and nonlinear synapses in the retina." Neuron **73**(4): 758-773.
Oesterle, J., C. Behrens, C. Schroder, T. Hermann, T. Euler, K. Franke, R. G. Smith, G. Zeck and P. Berens (2020). "Bayesian inference for biophysical neuron models enables stimulus optimization for retinal neuroprosthetics." Elife **9**.
Olshausen, B. A. and D. J. Field (1996). "Emergence of simple-cell receptive field properties by learning a sparse code for natural images." Nature **381**(6583): 607-609.
Panzeri, S., R. Senatore, M. A. Montemurro and R. S. Petersen (2007). "Correcting for the sampling bias problem in spike train information measures." J Neurophysiol **98**(3): 1064-1072.
Panzeri, S. and A. Treves (1996). "Analytical estimates of limited sampling biases in different information measures." Network **7**(1): 87-107.
Parsons, T. D. and P. Sterling (2003). "Synaptic ribbon. Conveyor belt or safety belt?" Neuron **37**(3): 379-382.
Pillow, J. W., L. Paninski, V. J. Uzzell, E. P. Simoncelli and E. J. Chichilnisky (2005). "Prediction and decoding of retinal ganglion cell responses with a probabilistic spiking model." J Neurosci **25**(47): 11003-11013.
Ramakrishnan, N. A., M. J. Drescher and D. G. Drescher (2012). "The SNARE complex in neuronal and sensory cells." Mol Cell Neurosci **50**(1): 58-69.
Redman, S. (1990). "Quantal analysis of synaptic potentials in neurons of the central nervous system." Physiol Rev **70**(1): 165-198.
Resnick, S. (1992). Adventures in stochastic processes. Basel, Switzerland, Birkhauser.
Ruderman, D. L. (1997). "Origins of scaling in natural images." Vision research **37**(23): 3385-3398.
Rudolph, S., M. C. Tsai, H. von Gersdorff and J. I. Wadiche (2015). "The ubiquitous nature of multivesicular release." Trends Neurosci **38**(7): 428-438.
Rutherford, M. A., N. M. Chapochnikov and T. Moser (2012). "Spike encoding of neurotransmitter release timing by spiral ganglion neurons of the cochlea." J Neurosci **32**(14): 4773-4789.

Sanes, J. R. and S. L. Zipursky (2010). "Design principles of insect and vertebrate visual systems." Neuron **66**(1): 15-36.
Scheer, N. and J. A. Campos-Ortega (1999). "Use of the Gal4-UAS technique for targeted gene expression in the zebrafish." Mechanisms of development **80**(2): 153-158.
Schmitz, F., A. Konigstorfer and T. C. Sudhof (2000). "RIBEYE, a component of synaptic ribbons: a protein's journey through evolution provides insight into synaptic ribbon function." Neuron **28**(3): 857-872.
Schroeder, C., B. James, L. Lagnado and P. Berens (2019). "Approximate bayesian inference for a mechanistic model of vesicle release from a ribbon synapse." Advances in neural information processing systems **33**.
Schwartz, O., J. W. Pillow, N. C. Rust and E. P. Simoncelli (2006). "Spike-triggered neural characterization." J Vis **6**(4): 484-507.
Schwartz, O. and E. P. Simoncelli (2001). "Natural signal statistics and sensory gain control." Nat Neurosci **4**(8): 819-825.
Sejnowski, T. J. (1995). "Pattern recognition. Time for a new neural code?" Nature **376**(6535): 21-22.
Sengupta, B., S. B. Laughlin and J. E. Niven (2014). "Consequences of converting graded to action potentials upon neural information coding and energy efficiency." PLoS Comput Biol **10**(1): e1003439.
Shadlen, M. N. and W. T. Newsome (1998). "The variable discharge of cortical neurons: implications for connectivity, computation, and information coding." J Neurosci **18**(10): 3870-3896.
Shannon, C. (1948). "A Mathematical Theory of Communication." Bell System Technical Journal **27**(3): 379-423.
Singer, J. H., L. Lassova, N. Vardi and J. S. Diamond (2004). "Coordinated multivesicular release at a mammalian ribbon synapse." Nat Neurosci **7**(8): 826-833.
So, P. T., C. Y. Dong, B. R. Masters and K. M. Berland (2000). "Two-photon excitation fluorescence microscopy." Annu Rev Biomed Eng **2**: 399-429.
Stein, R. B. (1965). "A Theoretical Analysis of Neuronal Variability." Biophys J **5**: 173-194.
Stevens, C. F. and Y. Wang (1995). "Facilitation and depression at single central synapses." Neuron **14**(4): 795-802.
Strong, S. P., R. R. de Ruyter van Steveninck, W. Bialek and R. Koberle (1998). "On the application of information theory to neural spike trains." Pac Symp Biocomput: 621-632.
Svoboda, K. and R. Yasuda (2006). "Principles of two-photon excitation microscopy and its applications to neuroscience." Neuron **50**(6): 823-839.
Temizer, I., J. C. Donovan, H. Baier and J. L. Semmelhack (2015). "A Visual Pathway for Looming-Evoked Escape in Larval Zebrafish." Curr Biol **25**(14): 1823-1834.
Theunissen, F. E. and J. P. Miller (1991). "Representation of sensory information in the cricket cercal sensory system. II. Information theoretic calculation of system accuracy and optimal tuning-curve widths of four primary interneurons." J Neurophysiol **66**(5): 1690-1703.
tom Dieck, S., W. D. Altrock, M. M. Kessels, B. Qualmann, H. Regus, D. Brauner, A. Fejtova, O. Bracko, E. D. Gundelfinger and J. H. Brandstatter (2005). "Molecular dissection of the photoreceptor ribbon synapse: physical interaction of Bassoon and RIBEYE is essential for the assembly of the ribbon complex." J Cell Biol **168**(5): 825-836.
Tong, G. and C. E. Jahr (1994). "Multivesicular release from excitatory synapses of cultured hippocampal neurons." Neuron **12**(1): 51-59.
Triller, A. and H. Korn (1982). "Transmission at a central inhibitory synapse. III. Ultrastructure of physiologically identified and stained terminals." J Neurophysiol **48**(3): 708-736.
Tuckwell, H. C. (1988). Introduction to Theoretical Neurobiology, Cambridge University Press.
Vaden, J. H., G. Banumurthy, E. S. Gusarevich, L. Overstreet-Wadiche and J. I. Wadiche (2019). "The readily-releasable pool dynamically regulates multivesicular release." Elife **8**.

Vere-Jones, D. (1966). "Simple stochastic models for the release of quanta of transmitter from a nerve terminal." Aust. J. Statist. 8 **8**: 53-63.
Vladimirov, N., C. Wang, B. Hockendorf, A. Pujala, M. Tanimoto, Y. Mu, C. T. Yang, J. D. Wittenbach, J. Freeman, S. Preibisch, M. Koyama, P. J. Keller and M. B. Ahrens (2018). "Brain-wide circuit interrogation at the cellular level guided by online analysis of neuronal function." Nat Methods **15**(12): 1117-1125.
Von Gersdorff, H. and G. Mathews (1994). "Dynamics of synaptic vesicle fusion and membrane retrieval in synaptic terminals." Nature **367**(6465): 735-739.
von Gersdorff, H., T. Sakaba, K. Berglund and M. Tachibana (1998). "Submillisecond kinetics of glutamate release from a sensory synapse." Neuron **21**(5): 1177-1188.
Wadiche, J. I. and C. E. Jahr (2001). "Multivesicular release at climbing fiber-Purkinje cell synapses." Neuron **32**(2): 301-313.
Wassle, H. and B. B. Boycott (1991). "Functional architecture of the mammalian retina." Physiol Rev **71**(2): 447-480.
White, K. A. and B. N. Kim (2021). "Quantifying neurotransmitter secretion at single-vesicle resolution using high-density complementary metal–oxide–semiconductor electrode array." Nature Communications **12**(1): 431.
Widrow, B., Y. Kim, D. Park and J. K. Perin (2019). Chapter 1 - Nature's Learning Rule: The Hebbian-LMS Algorithm. Artificial Intelligence in the Age of Neural Networks and Brain Computing. R. Kozma, C. Alippi, Y. Choe and F. C. Morabito, Academic Press**:** 1-30.
Williamson, R. S., M. Sahani and J. W. Pillow (2015). "The equivalence of information-theoretic and likelihood-based methods for neural dimensionality reduction." PLoS Comput Biol **11**(4): e1004141.
Wittig, J. H., Jr. and T. D. Parsons (2008). "Synaptic ribbon enables temporal precision of hair cell afferent synapse by increasing the number of readily releasable vesicles: a modeling study." J Neurophysiol **100**(4): 1724-1739.
Yeandle, S. and J. B. Spiegler (1973). "Light-evoked and spontaneous discrete waves in the ventral nerve photoreceptor of Limulus." J Gen Physiol **61**(5): 552-571.
Yusim, K., H. Parnas and L. A. Segel (2001). "One-vesicle hypothesis for neurotransmitter release: a possible molecular mechanism." Bull Math Biol **63**(6): 1025-1040.
Yuste, R. (2005). "Fluorescence microscopy today." Nat Methods **2**(12): 902-904.
Zeldenrust, F., P. Chameau and W. J. Wadman (2018). "Spike and burst coding in thalamocortical relay cells." PLoS Comput Biol **14**(2): e1005960.
Zeldenrust, F., P. J. Chameau and W. J. Wadman (2013). "Reliability of spike and burst firing in thalamocortical relay cells." J Comput Neurosci **35**(3): 317-334.
Zeldenrust, F., W. J. Wadman and B. Englitz (2018). "Neural Coding With Bursts-Current State and Future Perspectives." Front Comput Neurosci **12**: 48.
Zenisek, D., N. K. Horst, C. Merrifield, P. Sterling and G. Matthews (2004). "Visualizing synaptic ribbons in the living cell." J Neurosci **24**(44): 9752-9759.
Zheng, L., A. Nikolaev, T. J. Wardill, C. J. O'Kane, G. G. de Polavieja and M. Juusola (2009). "Network adaptation improves temporal representation of naturalistic stimuli in Drosophila eye: I dynamics." PLoS One **4**(1): e4307.
Zhou, M., J. Bear, P. A. Roberts, F. K. Janiak, J. Semmelhack, T. Yoshimatsu and T. Baden (2020). "Zebrafish Retinal Ganglion Cells Asymmetrically Encode Spectral and Temporal Information across Visual Space." Curr Biol **30**(15): 2927-2942 e2927.
Zimmermann, M. J. Y., N. E. Nevala, T. Yoshimatsu, D. Osorio, D. E. Nilsson, P. Berens and T. Baden (2018). "Zebrafish Differentially Process Color across Visual Space to Match Natural Scenes." Curr Biol **28**(13): 2018-2032 e2015.
Zipfel, W. R., R. M. Williams and W. W. Webb (2003). "Nonlinear magic: multiphoton microscopy in the biosciences." Nature Biotechnology **21**: 1369.

Zucker, R. S. (1973). "Changes in the statistics of transmitter release during facilitation." J Physiol **229**(3): 787-810.

Mathematical Appendix

0.1: Poisson Process and Exponential.

The times between arrivals of a Poisson Process are exponentially distributed.

Let $\{\boldsymbol{N(t)} : \boldsymbol{t > 0}\}$ be a Poisson Process with parameter λ. Define the random variable $\boldsymbol{X}$ as the time before an arrival of this Process. Note that here we assume an event has occurred at time zero – although this is not necessary due to the memoryless properties of the Poisson Process. Then these two situations are identical: (X>t) and (N(t)=0), both different way of saying 'there were no events before time t'. Thus,

$$P(X > t) = P(N(t) = 0) = \frac{\lambda^0 e^{-\lambda}}{0!} = e^{-\lambda}$$

Where the last term is by definition of a Poisson Process. Taking the complement of both sides yields:

$$P(X \leq t) = 1 - e^{-\lambda}$$

This is the cdf of the exponential distribution, so the proof is complete.

0.2: Poisson Process and Erlang.

The Time Between K arrivals to a Poisson Process is Erlang Distributed:

Here, we have already shown that if we have a PP, then the interarrivals times are exponentially distributed. Now we must only prove that the time before k occurrences of an exponential distribution is Erlang distributed. To do so, denote each exponential distribution as the random variable E_i. We seek to find the distribution of times before the kth arrival of PP. This is analogous to finding the distribution of the sum of k independent exponential random variables, that is:

$$g(z) = P(Z = E_1 + \ldots + E_k)$$

where z is the time before k arrivals of an exponential distribution.

One manner in which to prove the statement is by noting that the sum of random variables can be represented as the convolution of their densities. We could then find the solution for the k = 2 case, and then prove the statement by induction. However, there is a simpler method to prove the statement, that of the moment-generating function (mgf):

$$M_X(t) \equiv E[e^{tX}] = \int_{x \in X} e^{tX} f(x) dx$$

where f(x) is the pdf of the underlying distribution, and the equality holds due to LOTUS. While this might seem like an arbitrary transformation, it can be a powerful tool, paired with other properties. For one, the mgf allows for another analytical way to describe a distribution's moments. Secondly, it allows for a simpler method of analytically defining the sum of random variables: the mgf of a sum of independent random variables is simply the product of each variables mgf. This, paired with the fact that if two distributions have identical moments, then they are identical distributions,

allows for a simple proof of the statement. The mgf of an exponential distribution is $M_X(t) = \frac{\lambda}{\lambda - t}$. Thus, the time before k arrivals of an exponential distribution will have an mfg:

$$M_Z(t) = \prod_{i=1}^{k} M_X(t) = \prod_{i=1}^{k} \frac{\lambda}{\lambda - t} = \left(\frac{\lambda}{\lambda - t}\right)^k = \left(1 - \frac{t}{\lambda}\right)^{-k}$$

This is the mgf of the Erlang distribution with parameters λ and k, so the proof is complete.

0.3: Poisson Splitting

In Poisson Splitting, we note the following truth: Given a PP with parameter λ, we can stochastically split each event into k disjoint sub-types with time-dependent probability:

$$p_i(t) = P(Event\ at\ time\ t\ is\ of\ group\ i |\ event\ at\ time\ t)$$

Then the resulting distribution of events of each type is Poisson with parameter $\lambda_i(t) = \int_0^t \lambda p_i(s) ds$, and each of the Poisson Processes are independent of one another. Note that here the distribution of events is no longer independent from the distribution of event types due to the introduction of time-dependence. However, the statement is still correct as the two variables are conditionally independent - the probability of an event belonging to a group depends only upon the existence of the corresponding event, not any previous or future events or their classification. The proof for the time-dependent case is straightforward but requires some knowledge of measure theory, but informally it utilizes the concept of marked Poisson Processes and Poisson Random Measures. Here, a value in an additional measurable space (such as event amplitude) can be defined, mapped from the initial distribution of Poisson points. As long as the distribution of the number of events occurring over a measurable region is Poisson, and the distribution of events in time can also be measured, the resulting distribution is a Poisson Random Measure. See (Resnick 1992, Chiu, Stoyan et al. 2013) for more information or a formal proof.

While this might seem like some trivial fact of mathematics, the knowledge that any PP stochastically sub-divided (whether time-dependent or -independent) is still a PP allows for analytic calculation of the expected number of vesicles for each event type. If we then assign a number to each type (as in the case of describing the number of vesicles in a glutamatergic event), we can analytically compute the total number of vesicles released in a given time.

0.4: Poisson Merging

Poisson merging states that the sum of k independent PPs will be a PP with intensity functions the sum of the components. First note that since N(0)=0 for any PP, then their sum N(0)+…+N(0)=0, satisfying the first property of a PP. As the subprocesses have independent increments, then the resulting sum also has independent increments, satisfying the second property. To show the resulting distribution is Poisson we could either use convolution, or - like we did for the sum of exponentials, use the mgfs. Note that the mgf of the Poisson distribution is:

$$M_{Nsum}(t) = \prod_{i=1}^{k} e^{\lambda_i(e^t - 1)} = e^{(\sum_{i=1}^{k} \lambda_i)(e^t - 1)}$$

This is the mgf of a Poisson distribution with parameter lambda sum, fulfilling the last requirement. The proof is complete.

Note that the proof for the nonhomogeneous case can be analogously shown for nonhomogeneous PPs.